Published online 28 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D747–D750***
doi:10.1093/nar/gkl995

ArrayExpress—a public database of microarray experiments and gene expression profiles

H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne*, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans and A. Brazma

European Bioinformatics Institute, EMBL-EBI Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received September 20, 2006; Revised October 27, 2006; Accepted October 30, 2006

ABSTRACT

ArrayExpress is a public database for high throughput functional genomics data. ArrayExpress consists of two parts—the ArrayExpress Repository, which is a MIAME supportive public archive of microarray data, and the ArrayExpress Data Warehouse, which is a database of gene expression profiles selected from the repository and consistently re-annotated. Archived experiments can be queried by experiment attributes, such as keywords, species, array platform, authors, journals or accession numbers. Gene expression profiles can be queried by gene names and properties, such as Gene Ontology terms and gene expression profiles can be visualized. ArrayExpress is a rapidly growing database, currently it contains data from >50 000 hybridizations and >1 500 000 individual expression profiles. ArrayExpress supports community standards, including MIAME, MAGE-ML and more recently the proposal for a spreadsheet based data exchange format: MAGE-TAB.
Availability: www.ebi.ac.uk/arrayexpress.

INTRODUCTION

ArrayExpress (1) is a public database for storing and providing access to high throughput functional genomics data. ArrayExpress consists of two components specialized for distinct purposes—the ArrayExpress Repository of publicly available archived experimental data and the ArrayExpress Data Warehouse of gene expression profiles.

ArrayExpress REPOSITORY

The ArrayExpress repository is a MIAME compliant (2) primary archive containing the original data related to publications or generated by consortia. ArrayExpress is one of the three databases recommended by the MGED society (3) for depositions of publication related microarray data the other two being Gene Expression Omnibus (4) and CiBEX (5). ArrayExpress provides the means to store pre-publication data confidentially whilst allowing access to authorized users such as journal editors and referees. The data are made publicly available upon publication of the paper to which they relate. During the last 2 years the ArrayExpress Repository has grown 5-fold to over 50 000 hybridizations (September 2006) organized into 1650 different experiments. More than 90% of the experiments relate to gene expression profiling studies, the remainder are array based chromatin immunoprecipitation or comparative genomics experiments. Over 200 different organisms are represented, the largest contributors being human, mouse, Arabidopsis, yeast and rat.

A new ArrayExpress experiment browse and query interface (Figures 1 and 2) was released in 2006. It allows the user to browse the entire content of the database in a summary view or query public datasets using free text and displays the query results in a summary view of up to 500 experiments per page which can be sorted by name, accession number and load date, and filtered by array design, species, date or availability of raw and processed data. For example, the user can query for all experiments containing the phrase 'bone marrow' in their description, or retrieve experiments by accession numbers (e.g. E-TABM-102), publication details, array design names and journal names.

Each row in the summary view can be expanded to a detailed view including experiment description and publication references. Queries can also be exported to spreadsheets and saved. There are links to the original data files where filled or empty icons indicate the presence or absence of raw or normalized data. Each experiment has a link to spreadsheets and graphs describing the sample properties and experiment design. Linking to the 'Detailed data retrieval page' allows for selection of particular Quantitation Types (e.g. signal or log ratio) from multiple data files for specific conditions. These can be exported into a single data matrix for download, or uploaded to Expression profiler, an online data analysis tool for further processing. An advanced

*To whom correspondence should be addressed. Tel: +44 1223 494 681; Fax: +44 1223 494 468; Email: farne@ebi.ac.uk

query interface is also provided http://www.ebi.ac.uk/aerep for password-protected access to private data and complex queries, e.g. all experiments performed on a specific array and complex combinations of species/experiments.

ArrayExpress DATA WAREHOUSE

The ArrayExpress Data Warehouse currently contains 1 500 000 gene expression profiles from >2500 hybridizations. s. It is currently populated with gene expression data from ArrayExpress. Selection for inclusion is on the basis of MIAME compliance, a curator's assessment of quality of annotation of the samples and arrays. For example, probe sequences must be available for sequence based matching or database identifiers that can be matched to Uniprot (6) should be present for the majority of spots on an array.

The query interface to the ArrayExpress Data Warehouse (Figures 3 and 4) allows the user to find gene expression profiles by gene name or identifier, database accession number (e.g. UniProt) or other annotation such as Gene Ontology terms. If a query returns more than one gene (for instance, a query for BRCA will match BRCA1 and BRAC2), then a list of matches is provided and the user can view the properties of these genes before selecting appropriate ones. The default view is the gene identifier, synonyms, Gene Ontology terms and links back to source databases. Once the user has selected one or more genes, all the experiments containing selected genes stored in the data warehouse are returned as a list with thumbnail images of their expression profiles. The user can then order this list by different statistical criteria quantifying the 'relevance' of these experiments for the selected genes, and can zoom in by clicking on these images, and view more information. For instance, one can investigate whether the expression of a selected gene changes with experimental variables such as disease states or cell types. Selecting genes with similar expression profiles, and the ability to export numerical values from the data warehouse is currently under development.

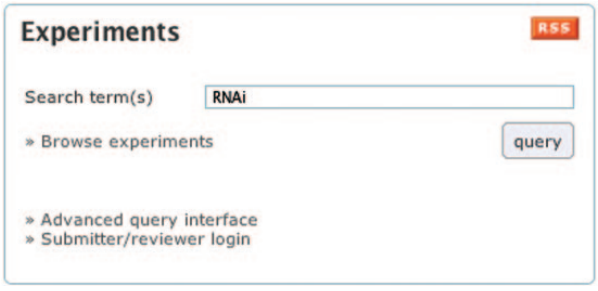

Figure 1. ArrayExpress experiment query form. Queries on experiment properties: organisms, author's names, array types or accession numbers are supported.

BULK DOWNLOADS

The ArrayExpress data can be downloaded from ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/. Data files are made available as either tab-delimited text files for two-color experiments or native CEL format files (Affymetrix) within zip archives. You can also download a graphical representation of each experiment, a spreadsheet showing the sample annotation and a MAGE-ML format file from the ftp site.

ANALYZING MICROARRAY DATA IN EXPRESSION PROFILER

Expression Profiler is an online microarray-data analysis tool that can be used either to analyze data retrieved from ArrayExpress or to analyze data uploaded from any other source, such as the user's own local private data (7). The user can import data to Expression Profiler from the ArrayExpress advanced query interface by selecting 'Data Export to Expression Profiler', or can download zip archives from the ArrayExpress browse interface and then upload them into Expression Profiler for normalization and analysis. Expression Profiler provides distinct, chainable components for clustering, pattern discovery, statistical analysis, machine-learning algorithms and visualization. All of Expression Profiler's components can be accessed using SOAP-based web services and can be incorporated into bioinformatics workflows using Taverna (8).

DATA SUBMISSION AND CURATION

Data can be submitted to the ArrayExpress repository by one of three routes.

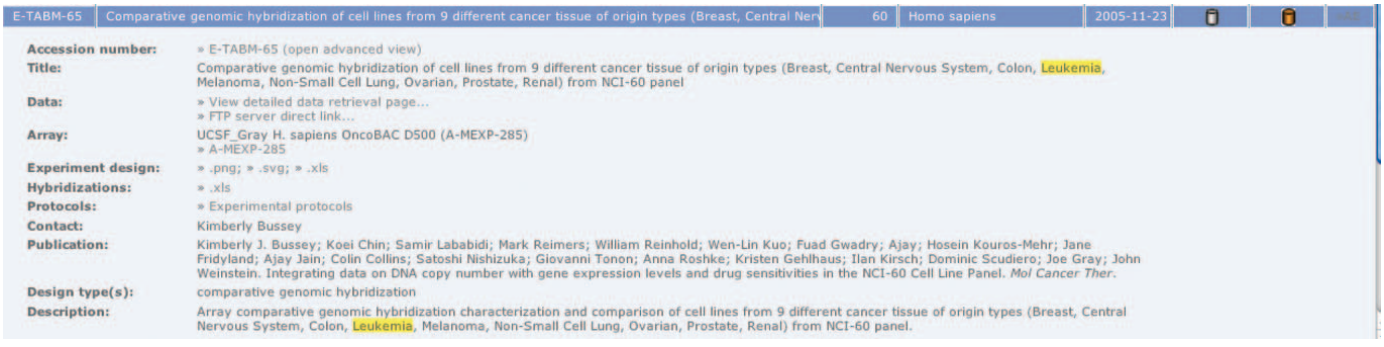

Figure 2. A detailed view of an experiment from the repository. Strings matching the query terms are highlighted in yellow.

Web based submissions

For submissions of up to ~20 hybridizations the online submission tool MIAMExpress is recommended (http://www.ebi.ac.uk/miamexpress/). A batch-loader for larger experiments via this route is currently being tested and will be released in 2006.

Spreadsheet submissions

Experiments of all types and sizes can be submitted as spreadsheets, http://www.ebi.ac.uk/cgi-bin/microarray/tab2mage.cgi. A new template generation system provides a user with a spreadsheet template based on technology type, species and experiment type. The user then downloads the spreadsheet, completes it, and uploads it with the data files. The spreadsheet uses a simple 'one-row-per-channel' model and is suitable for Affymetrix, two-color experiments and Nimblegen technology, ChIP-chip and CGH. Our experience with developing a spreadsheet based submission tool has led to a community developed tab-delimited data exchange format MAGE-TAB (9) which will be supported in 2007.

Submissions from external microarray databases

Laboratories that have a local database can develop automated data export using the MAGE-ML or MAGE-TAB formats, which can be submitted to ArrayExpress directly. The ArrayExpress curators work with external databases when setting up a data submission pipeline to ensure that data are MIAME compliant and well formatted. Once a pipeline is established, the submissions are curated at the source database and monitored by ArrayExpress curators. MAGE-ML based pipelines have been established from 20 external databases, manufacturers or tools.

Figure 3. Gene expression profiles query form. Queries on gene properties: names, accession numbers, synonyms and gene ontology terms are supported.

Curation

The ArrayExpress curation team processes each submission of experimental data before it is loaded into the repository. Submissions are checked for elements of MIAME compliance, in particular, the presence of raw and processed data, accuracy and completeness of biological information provided, including the presence of experimental factors and their values in the sample annotation. Data consistency is also checked, e.g. that submitted data files match the specified array designs and that data files are complete and uncorrupted. ArrayExpress has been requested by journals to provide a MIAME assessment service (10). This service is available on request from miamexpress@ebi.ac.uk at present, but all legacy data will have a MIAME score computed which will be displayed in the ArrayExpress user interface by the end of 2006 and will also be available to reviewers viewing data supporting publications. The MIAME scoring is currently used internally to select data for the Data Warehouse, and the scoring system is described in (10).

DATA WAREHOUSE CURATION

Ensembl (11) and UniProt (8) are used both as a source of updated and additional annotation for arrays and both provide access to gene expression data in the data warehouse from their own databases. In Ensembl this is achieved via a DAS track and since August 2006 in Uniprot (8) via database cross-references.

There are two distinct methodologies for re-annotation and an automated pipeline is in place for each

(i) Where the reporter sequences used on the array are public and are matched on one of the Ensembl genome builds Ensembl annotation is used. The process of mapping sequences for array features (probes) to genes is described in the Ensembl user documentation (http://www.ensembl.org/info/data/docs/microarray_probe_set_mapping.html).
(ii) Where no Reporter sequences are publicly available, Reporters are not mapped, or the species is not included in Ensembl, identifier matches in the UniProt database are used to acquire additional annotation.

As Ensembl has a frequent release cycle and annotation changes and probes can be re-mapped during this process we have designed an annotation update pipeline. An internal tracking database is used to store information on Ensembl annotation for probes on a weekly basis. When new

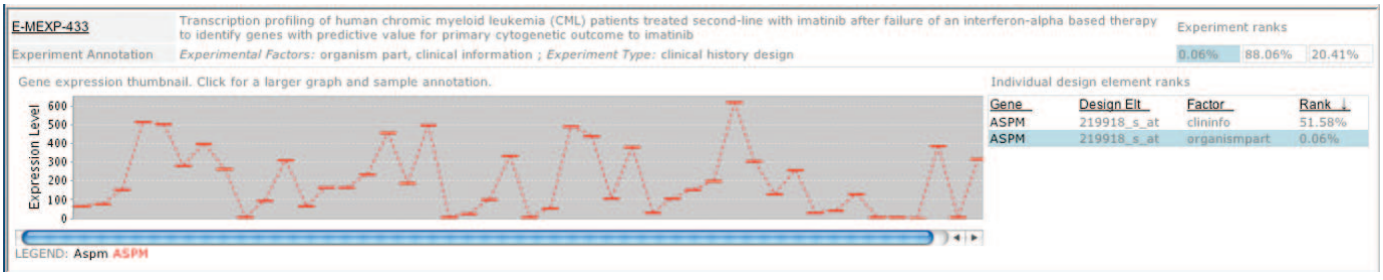

Figure 4. A view of gene expression for gene 'aspm' within a single experiment in the data warehouse.

annotation is detected old annotation is deleted from the data warehouse for probes and updated annotation extracted from Ensembl is inserted. Our version tracking system allows us to check statistics for the probes on specific arrays. For example, the percentage of probes from a given array that fail to map to the genome is monitored and can be checked for anomalies between Ensembl releases for each array design.

Experiments are identified as being suitable for inclusion in the Data Warehouse at the point of submission by matching the following criteria: the experiments must use an array that can be re-annotated, have normalized and processed data and be gene expression experiments that are MIAME compliant. The data warehouse provides an environment to update the annotation for arrays; however, the original array annotation supporting the publication is maintained in the repository in the state that it was submitted. Submitters are therefore not permitted to re-annotate arrays linked to published data as this may affect conclusions drawn from these data reported in the paper and this is inconsistent with the archival role of the ArrayExpress repository.

FUTURE

The ArrayExpress database has been online since 2002 and has not only grown from a few hundred to >50 000 hybridizations, but also has undergone major software and usability improvements. This is a continuous process, and among the immediate tasks are implementing more powerful query mechanisms to deal with providing simple access to large volumes of data, providing webservices, and deeper integration with the data analysis, mining and visualization tool Expression Profiler. We are constantly improving the ease of submission to ArrayExpress, and we expect that the new generation of submission tools, together with a wide adoption of MAGE-TAB format will further increase the ease of submission. MAGE-TAB will be supported at ArrayExpress in 2007.

A prototype XML interface is available which exposes Experiment level information plus sample annotation. Programmatic interfaces are under development and we invite users to submit their use cases for such systems. MIAME scores will be computed for each experiment and displayed in the user interface for all legacy data as well as for new submissions.

In the future, as the community develops data quality metrics, we will include these as criteria for inclusion of data in the Data Warehouse and we will display these for experiment submissions.

Work is under way to integrate the most commonly used Bioconductor modules into Expression Profiler during the next 12 months, thereby providing a web interface to these. Finally, we have begun a gene atlas project where data in the public domain are re-normalized consistently in-house and will be loaded into the Data Warehouse as supersets of data.

URLs

ArrayExpress home page and query interface: http://www.ebi.ac.uk/arrayexpress/; MIAMExpress data submission tool: http://www.ebi.ac.uk/miamexpress/; Tab2mage submission tool: http://www.ebi.ac.uk/cgi-bin/microarray/tab2mage.cgi.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Patrick Kemmeren, Susanna-Assunta Sansone, Philippe Rocca-Serra, Catherine Brooksbank, Petteri Jokkinen and the EBI systems group and the ArrayExpress Scientific Advisory Board. EMBL, the European Commission grants: TEMBLOR, FELICS and MUGEN support ArrayExpress Development. Funding to pay the Open Access publication charges for this article was provided by the FELICS grant from the European Commission.

Conflict of interest statement. None declared.

REFERENCES

1. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Lara,G.G., Holloway,E., Kapushesky,M. *et al*. (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **33**, D553–D555.
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. *Nature Genet*., **29**, 365–371.
3. Ball,C.A., Brazma,A., Causton,H., Chervitz,S., Edgar,R., Hingamp,P., Matese,J.C., Parkinson,H.E., Quackenbush,J., Ringwald,M. *et al*. (2004) Submission of microarray data to public repositories. *PLoS Biol*., **2**, E317.
4. 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.
5. Ikeo,K., Ishi-I,J., Tamura,T., Gojobori,T. and Tateno,Y. (2003) CIBEX: center for information biology gene expression database. *C. R. Biol*., **326**, 1079–1082.
6. 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*., **134**, D187–D191.
7. Kapushesky,M., Kemmeren,P., Culhane,A.C., Durinck,S., Ihmels,J., Korner,C., Kull,M., Torrente,A., Sarkans,U., Vilo,J. *et al*. (2006) Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucleic Acids Res*., **32**, W465–W470.
8. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*., **34**, W729–W732.
9. Rayner,T., Rocca-Serra,P., Spellman,P.T., Causton,H.C., Farne,A., Holloway,E., Liu,J., Maier,D.S., Miller,M., Petersen,K. *et al*. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.
10. Brazma,A. and Parkinson,H. (2006) ArrayExpress service for reviewers/editors of DNA microarray papers. *Nature Biotech*., **24**, 1321–1322.
11. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.