D1002–D1004 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 10 November 2010*
doi:10.1093/nar/gkq1040

ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments

Helen Parkinson[1,*], Ugis Sarkans[1], Nikolay Kolesnikov[1], Niran Abeygunawardena[1], Tony Burdett[1], Miroslaw Dylag[1], Ibrahim Emam[1], Anna Farne[1], Emma Hastings[1], Ele Holloway[1], Natalja Kurbatova[1], Margus Lukk[2], James Malone[1], Roby Mani[1], Ekaterina Pilicheva[1], Gabriella Rustici[1], Anjan Sharma[1], Eleanor Williams[1], Tomasz Adamusiak[1], Marco Brandizi[1], Nataliya Sklyar[1] and Alvis Brazma[1]

[1]Functional Genomics Team, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD and [2]Regulatory Systems Biology Laboratory, Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

Received September 14, 2010; Revised and Accepted October 11, 2010

ABSTRACT

The ArrayExpress Archive (http://www.ebi.ac.uk/arrayexpress) is one of the three international public repositories of functional genomics data supporting publications. It includes data generated by sequencing or array-based technologies. Data are submitted by users and imported directly from the NCBI Gene Expression Omnibus. The ArrayExpress Archive is closely integrated with the Gene Expression Atlas and the sequence databases at the European Bioinformatics Institute. Advanced queries provided via ontology enabled interfaces include queries based on technology and sample attributes such as disease, cell types and anatomy.

INTRODUCTION

The ArrayExpress Archive of Functional Genomics Data is one of the major international repositories for functional genomics high-throughput data. Since 2003 (1), the database has grown to ~15 000 experiments comprised of ~425 000 assays. During this period, the technology used to generate functional genomics data has changed from microarray-based experiments to high-throughput sequencing. To address this, we have developed and integrated submissions of high-throughput sequencing data with the European Genome-phenome Archive (EGA) (Lappalainen,I. *et al.*, submitted) and the European Nucleotide Archive (ENA) (2). Other important developments are the inclusion of all Gene Expression Omnibus (GEO) array-based data and a new data exchange agreement with the GEO (3) for high-throughput sequencing data, a new advanced query capability supporting ontology-based queries over the entire Archive contents. The European Bioinformatics Institute's Gene Expression Atlas (GXA) (4) is now a separate resource from the Archive and is linked from the ArrayExpress Graphical User Interface.

Support for high-throughput sequencing data

Adjusting ArrayExpress to accept and display high-throughput sequencing experiments alongside existing array data is one of the major recent developments. We have worked closely with other resources at European Bioinformatics Institute, specifically the ENA and EGA, who archive short-read data for multi-species and potential human identifiable data, respectively. As outlined in MINSEQE guidelines (Minimum Information about a high throughput Sequencing Experiment, http://www.mged.org/minseqe), the provision of raw sequence data is insufficient to describe comparative experiments such as RNA-Seq; metadata describing the experimental conditions and processed data are necessary to interpret the experiment. There are parallels to the provision of metadata for microarray-based experiments (in addition to the raw data files, e.g. CEL files); therefore, the MAGE-TAB (5) data representation format is both an appropriate and an easy to use format for describing these experiments.

Submission of high-throughput sequencing data are now supported by the MAGE-TAB template generation system (6). This allows users to generate and complete a taxon-specific tab-delimited template which describes their experiment and to supply related data files by FTP or Aspera. Where raw data are available these are integrated

*To whom correspondence should be addressed. Tel: +44 1223 494672; Fax: +44 1223 494468; Email: parkinson@ebi.ac.uk

into the ENA at the point of submission, and both ArrayExpress experiment accessions and ENA identifiers for sequences are returned to the user by ArrayExpress curators. Exceptions to this process are submissions with human data which are potentially identifiable, e.g. sequence of human patients. These data are submitted direct to the EGA in MAGE-TAB format, raw data are retained by the EGA and summary-level data which meet ethical requirements for release are released to ArrayExpress.

Sequencing-based experiments in ArrayExpress now have clickable links from the user interface to the ENA sequence archive to raw data files, and links are also provided in the MAGE-TAB. Work is in progress to develop an automated BioConductor (7) package to identify, extract and reprocess RNA-Seq data for inclusion in the GXA.

Advanced queries and ontology-driven searches

ArrayExpress provides rich metadata for samples and experiments, these are typically provided as free-text name value pairs, e.g. disease state, invasive ductal carcinoma. To enable semantic queries (for instance, to find all cancer-related data sets even if they were not annotated as 'cancer', but e.g. 'leukemia'), we have developed open source software that allows for query expansion based on the Experiment Factor Ontology (EFO) (8). EFO is a data-driven application ontology developed to describe the sample attributes and experimental variables in functional genomics data sets. The new advanced query syntax allows logical, range and ontology-supported queries. For example, 'retrieve all experiments where one or more samples is annotated as cancer, or a subtype of cancer' returns 21 083 assays, without the ontology support and 49 729 assays using subsumption queries for known subtypes of cancer. The query results are visualized with yellow matching original input, green matching synonyms and red matching child terms. The ontology is visualized as a tree on query and users are provided with autocomplete options based on its content. Additionally, the interface has been modified so that experiments can be queried by assay types (array/high-throughput sequencing), source (GEO/ArrayExpress) and molecule (DNA/RNA).

Integration of GEO data

ArrayExpress has been importing selected GEO (3) Data sets (GDS) in order to provide unified queries across public data and for integration with European Bioinformatics Institute databases such as Ensembl (9). All GEO data with GDS and GSE prefixes are now being imported into ArrayExpress. To date more than 12 000 GEO-derived experiments and associated array designs are available, import of all GEO data will be complete by the end of 2010. Selected GDS are re-annotated, subjected to quality control and integrated into the GXA. A data exchange agreement between GEO and ArrayExpress is now in place for high-throughput sequencing data and all HTP sequencing data submitted to GEO are present in ArrayExpress.

Supporting software

Software developed tools for curation processes are publicly available; recent software releases include a lexical mapping application, Zooma (http://zooma.sf.net), the Ontocat ontology query service (http://ontocat.sf.net), a canonical MAGE-TAB parser (http://limpopo.sf.net), MAGE-TAB format conversion tools (http://tab2mage.sf.net) and a MAGE-TAB importer for Bioconductor (10).

Future developments

ArrayExpress will be closely integrated with a new BioSample Database at the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/biosamples). This database will store the sample descriptions for all the samples referenced by any of the databases. Samples can be pre-submitted and will be linked to EBI databases where related data exist. For example, 1000 genomes, Coriell cell lines or HapMap samples have records in the ENA, EGA and ArrayExpress. This new resource is being developed in conjunction with the NCBI and data exchange is planned.

The replacement of existing MAGE-OM centric architecture (11) with MAGE-TAB-based infrastructure is ongoing and data migration is underway. This effort will significantly simplify all internal data management tasks and will benefit the users in improved data load times, faster issuing of accession numbers, faster data exchange with GEO and improved query interfaces. The existing browse user interface will be maintained, as will current programmatic access and FTP site structure to ensure minimum disruption for users.

FUNDING

European Molecular Biology Laboratory, the European Commission (SLING grant agreement number 226073, Gen2Phen grant agreement number 200754); US National Institutes of Health (the National Human Genome Research Institute, the National Institute of Biomedical Imaging and Bioengineering and the National Cancer Institute) (grant number P41 HG003619). Funding for open access charge: European Molecular Biology Laboratory Functional Genomics Team Budget.

Conflict of interest statement. None declared.

REFERENCES

1. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
2. Cochrane,G., Akhtar,R., Bonfield,J., Bower,L., Demiralp,F., Faruque,N., Gibson,R., Hoad,G., Hubbard,T., Hunter,C. *et al.* (2009) Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, **37**, D19–D25.
3. 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.
4. Kapushesky,M., Emam,I., Holloway,E., Kurnosov,P., Zorin,A., Malone,J., Rustici,G., Williams,E., Parkinson,H. and Brazma,A. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.*, **38**, D690–D698.
5. Rayner,T., Rocca-Serra,P., Spellman,P.T., Causton,H.C., Farne,A., Holloway,E., Liu,J., Maier,D.S., Miller,M., Petersen. *et al*. A simple spreadsheet-based, MIAME-supportive format for microarray data. *BMC Bioinformatics*., **7**, 489.
6. Rayner,T.F., Rezwan,F.I., Lukk,M., Bradley,X.Z., Farne,A., Holloway,E., Malone,J., Williams,E. and Parkinson,H. (2009) MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB. *Bioinformatics*, **25**, 279–280.
7. Gentleman,R., Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds), (2005) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York.
8. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
9. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
10. Kauffmann,A., Rayner,T.F., Parkinson,H., Kapushesky,M., Lukk,M., Brazma,A. and Huber,W. (2009) Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics*, **25**, 2092–2094.
11. Spellman,P.T., Miller,M., Stewart,J., Troup,C., Sarkans,U., Chervitz,S., Bernhart,D., Sherlock,G., Ball,C., Lepage,M. *et al*. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, **3**, RESEARCH0046.