Published online 27 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D987–D990***
doi:10.1093/nar/gks1174

ArrayExpress update—trends in database growth and links to data analysis tools

Gabriella Rustici[1,*], Nikolay Kolesnikov[1], Marco Brandizi[1], Tony Burdett[1], Miroslaw Dylag[1], Ibrahim Emam[1], Anna Farne[2], Emma Hastings[1], Jon Ison[1], Maria Keays[1], Natalja Kurbatova[1], James Malone[1], Roby Mani[1], Annalisa Mupo[2], Rui Pedro Pereira[1], Ekaterina Pilicheva[1], Johan Rung[1], Anjan Sharma[1], Y. Amy Tang[1], Tobias Ternent[1], Andrew Tikhonov[1], Danielle Welter[1], Eleanor Williams[1], Alvis Brazma[1], Helen Parkinson[1] and Ugis Sarkans[1]

[1]Functional Genomics Team, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton CB10 1SD and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

Received October 19, 2012; Revised October 26, 2012; Accepted October 28, 2012

ABSTRACT

The ArrayExpress Archive of Functional Genomics Data (http://www.ebi.ac.uk/arrayexpress) is one of three international functional genomics public data repositories, alongside the Gene Expression Omnibus at NCBI and the DDBJ Omics Archive, supporting peer-reviewed publications. It accepts data generated by sequencing or array-based technologies and currently contains data from almost a million assays, from over 30 000 experiments. The proportion of sequencing-based submissions has grown significantly over the last 2 years and has reached, in 2012, 15% of all new data. All data are available from ArrayExpress in MAGE-TAB format, which allows robust linking to data analysis and visualization tools, including Bioconductor and GenomeSpace. Additionally, R objects, for microarray data, and binary alignment format files, for sequencing data, have been generated for a significant proportion of ArrayExpress data.

INTRODUCTION

The ArrayExpress Archive of Functional Genomics Data (1) is one of the major international repositories for functional genomics high throughput data, supporting publications as well as various data generating consortia. It stores functional genomics data derived from high throughput sequencing (HTS) and microarray-based experiments. Users come to ArrayExpress to (i) find functional genomics experiments that might be relevant to their research; (ii) retrieve information describing these experiments and the data associated with them; (iii) retrieve data for including in their own local data warehouses or added value databases; and (iv) submit their own data supporting a peer-reviewed publication.

Once submitted, data may be kept in ArrayExpress as private for a limited period of time, typically during the peer-review process of the related publication. Upon submission, an accession number is assigned to it and access to the data is restricted to providers/reviewers via a login system. The submitter specifies the release date and the data becomes public either when the accession number associated with the data is cited in a publication or at the set release date, whichever comes first.

All submissions are automatically checked for compliance to the Minimum Information About a Microarray Experiments (MIAME) (2) or Minimum Information about Sequencing Experiments (MINSEQE – http://www.fged.org/projects/minseqe/) guidelines, for microarray and sequencing-based experiments, respectively. The MIAME/MINSEQE scores associated with an experiment are displayed in the ArrayExpress interface and provided to submitters.

In addition to the data submitted directly to ArrayExpress, data from the Gene Expression Omnibus (GEO) (3) are imported to provide users with a single access to most of the functional genomics data available in the public domain. All data are organized, and available for download, in a structured and standardized format, MAGE-TAB (4), which also facilitates linking to open source analysis environments such as Bioconductor (5) and GenomeSpace (http://www.genomespace.org). A format conversion tool, from GEO SOFT to MAGE-TAB (6), is run on all GEO HTS and microarray data. The conversion is successful in 83% of cases; there are various reasons why this conversion may fail, including

*To whom correspondence should be addressed. Tel: +44 1223 492539; Fax: +44 1223 494468; Email: gabry@ebi.ac.uk

failure to parse SOFT files correctly or failure to retrieve the associated data files and we are constantly working with GEO to increase the success rate. All HTS data are exchanged with GEO and a data sharing agreement with the DDBJ Omics Archive is also in place (7).

For all experiments, the column labels describing the sample (e.g. disease) and its characteristics (e.g. type II diabetes) are mapped to the EBI's Experimental Factor Ontology (EFO) (8) and the data loaded into ArrayExpress. This allows consistent query results to be returned from direct submissions as well as imported data. As data are curated for Gene Expression Atlas use (9), they are reloaded into ArrayExpress with enriched annotation.

The ArrayExpress user interface allows users to search for experiments of interest by keywords and ontology terms, which enable semantically driven searches of the experimental metadata; for instance searching with the EFO term 'cancer' will also find experiments investigating 'leukemia' even if 'cancer' is not mentioned explicitly. Both US and UK spelling is supported.

DATA GROWTH TO A MILLION ASSAYS

Over the last 2 years, the database content has grown from 13 000 experiments and 370 000 assays, to over 30 000 experiments and almost a million assays. Approximately 20% of the data were submitted directly to ArrayExpress; the rest are imported from GEO weekly.

Although HTS-based experiments account for only 6% of the entire database content, the proportion of new HTS submissions has been growing exponentially over the last few years, from 2% in 2009 to 6% in 2010, 7% in 2011 and 15% in 2012. Nevertheless, the total number of assays associated with HTS-based experiments is still only 3%, reflecting the fact that HTS experiments are typically smaller than microarray-based experiments. If we look at a breakdown of the HTS data by application, 50% of the experiments used RNA-seq only, 32% ChIP-seq only and the remaining experiments either utilized more than one application or used DNA-seq for genotyping, copy number variation detection or methylation profiling.

For HTS data, ArrayExpress stores processed data and metadata describing the sample properties and the experimental design, including experimental variables and protocols, whereas raw sequence data are stored in the European nucleotide archive (ENA) (10) and linked from ArrayExpress. For datasets that require controlled access, the raw sequence data are stored in, and should be submitted directly to, the European Genome-phenome Archive (EGA – www.ebi.ac.uk/ega).

LINKS TO DATA ANALYSIS TOOLS

Approximately 50 GB of data are downloaded every day from ArrayExpress, by an average of 1000 different users. To simplify the interface between ArrayExpress and analytical platforms, we are now providing links to popular analytical tools such as Bioconductor and GenePattern (11), as well as developing robust internal pipelines for HTS data processing.

To facilitate loading microarray data from ArrayExpress into Bioconductor, we have pre-generated R objects for 16 250 out of 25 000 gene expression microarray experiments with raw data files available. A revised version of the Bioconductor package ArrayExpress (12) is used with default parameters. The package has been updated to support popular data formats including Affymetrix and Agilent. More than 85% of Affymetrix data in the repository have downloadable R objects. Older submissions, other technologies and experiments with only processed data available can still be loaded in R, but require user-specified settings for the package to recognize the data format, so loading must be supervised by a user. All pregenerated R objects are now available through the ArrayExpress interface and can be easily loaded into Bioconductor for downstream analysis. More R objects will be created for experiments in ArrayExpress as more data arrive, and the R package will be maintained and extended for this purpose.

Direct links are now provided to GenomeSpace (http://www.genomespace.org), a data analysis environment that makes it possible for users to move data smoothly between popular bioinformatics tools. From ArrayExpress, the user can, with a single click, load a dataset into GenomeSpace, provided that he/she has a registered account with GenomeSpace. Once logged in, the user will be able to utilize the data analysis tools available through GenomeSpace, including GenePattern, Galaxy (13) and Cytoscape (14), to perform data analysis.

For HTS data, the Bioconductor package ArrayExpressHTS (15) and the R-workbench (http://www.ebi.ac.uk/Tools/rcloud/) are used to generate binary alignment (BAM) format files (16). BAM files contain sequence alignment data and can be displayed using the Ensembl genome browser (17), through a direct link from ArrayExpress. So far approximately 1200 BAM files are available for 125 RNA-seq experiments, for 14 different species, with over half of these data studying human and a quarter mouse. The BAM file generation has been done for experiments for which: (i) the sample–data relationship information is available and contains details such as the library strategy and the experiment type (i.e. RNA-seq); (ii) the raw sequence reads (FASTQ files) are deposited in ENA and a valid link to the ENA entry is present; and (iii) the annotation for the reference genome is available in Ensembl.

In addition, 3000 datasets from ArrayExpress have been analysed and the results of this analysis are presented through the Gene Expression Atlas (9), a separate EBI database, which helps users to (i) find out whether the expression of a gene (or a group of genes with a common gene attribute, e.g. GO term) change(s) across all the experiments or (ii) discover which genes are differentially expressed in a particular biological condition of interest.

CONTINUOUS USER INTERFACE IMPROVEMENTS

The ArrayExpress user interface has been continuously improved since the repository was established in

Experiment E-MTAB-513

RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map) (19 samples)

	Sample Characteristics					*Factor Values*		*Links to Data*
Source Name	**Organism**	**Age**	**(unit)**	**OrganismPart**	**Sex**	**ORGANISMPART**	**LIBRARYPREP**	**ENA**
HCT20142	Homo sapiens	60	years	kidney	female	kidney	mRNA-Seq	
HCT20142	Homo sapiens	60	years	kidney	female	kidney	mRNA-Seq	
HCT20142	Homo sapiens	60	years	kidney	female	kidney	mRNA-Seq	
HCT20143	Homo sapiens	77	years	heart	male	heart	mRNA-Seq	
HCT20143	Homo sapiens	77	years	heart	male	heart	mRNA-Seq	
HCT20143	Homo sapiens	77	years	heart	male	heart	mRNA-Seq	
HCT20144	Homo sapiens	37	years	liver	male	liver	mRNA-Seq	
HCT20144	Homo sapiens	37	years	liver	male	liver	mRNA-Seq	
HCT20144	Homo sapiens	37	years	liver	male	liver	mRNA-Seq	
HCT20145	Homo sapiens	65	years	lung	male	lung	mRNA-Seq	
HCT20145	Homo sapiens	65	years	lung	male	lung	mRNA-Seq	
HCT20145	Homo sapiens	65	years	lung	male	lung	mRNA-Seq	
HCT20146	Homo sapiens	86	years	lymph node	female	lymph node	mRNA-Seq	
HCT20146	Homo sapiens	86	years	lymph node	female	lymph node	mRNA-Seq	
HCT20146	Homo sapiens	86	years	lymph node	female	lymph node	mRNA-Seq	
HCT20147	Homo sapiens	73	years	prostate	male	prostate	mRNA-Seq	
HCT20147	Homo sapiens	73	years	prostate	male	prostate	mRNA-Seq	
HCT20147	Homo sapiens	73	years	prostate	male	prostate	mRNA-Seq	
HCT20148	Homo sapiens	77	years	skeletal muscle	male	skeletal muscle	mRNA-Seq	
HCT20148	Homo sapiens	77	years	skeletal muscle	male	skeletal muscle	mRNA-Seq	
HCT20148	Homo sapiens	77	years	skeletal muscle	male	skeletal muscle	mRNA-Seq	

Figure 1. Sample–data relationship viewer for Experiment E-MTAB-513. This view provides information on sample characteristics and experimental variables that are fundamental to understand the results obtained in the experiment. Generally, each row corresponds to a sample. Columns include sample characteristics and their relationship to the resulting data files, providing a quick view over the structure of the experiment and the biological questions that the authors addressed. The last column provides links to raw sequence data files available in ENA, and BAM files that can be visualized in the Ensembl genome browser.

2003 (18). Recent additions include the sample–data relationship viewer (Figure 1), which provides an overview of all samples used in an experiment and their characteristics, the experimental variables (factors) investigated and the data files associated with each sample.

Other improvements include (i) improved array designs browsing and querying for; (ii) specific features for HTS data display; (iii) better organization of the species drop-down filter, and (iv) improved performance for retrieving and visualizing large experiments.

The ArrayExpress user documentation has recently been updated and several online courses, covering how to search, interpret and submit data to ArrayExpress, can be found on the EBI e-Learning portal, Train online (http://www.ebi.ac.uk/training/online/).

FUTURE DEVELOPMENTS

We are currently developing a new submission tool, optimized for supporting HTS data submissions; this new tool is based on the community developed annotation tool Annotare (19) and will be released in 2013.

Like all other major EBI data resources, ArrayExpress is working toward deeper integration in the overall EBI infrastructure, in particular with the BioSample Database (20), the Gene Expression Atlas and the sequence databases ENA, EGA and Ensembl. We will continue this integration effort to ensure that our users can obtain a systems level view of the data stored at EBI by easily navigating through our resources.

FUNDING

ArrayExpress and related activities are supported by member states of the European Molecular Biology Laboratory; European Commission: ENGAGE [201413], EurocanPlatform [260791], GEUVADIS [261123], SLING [226073], SYBARIS [242220], and Gen2Phen [200754]; US National Institutes of Health (the National Human Genome Research Institute, National Institute of Biomedical Imaging and Bioengineering and the National Cancer Institute) [P41 HG003619]; National Center for Biomedical Ontology, one of the National Centers for Biomedical Computing supported by the National Human Genome Research Institute, the National Heart, Lung, and Blood Institute, and National Institutes of Health Common Fund [U54-HG004028]; National Science Foundation Award Number [1127112]. Funding for open access charge: EMBL Members states.

Conflict of interest statement. None declared.

REFERENCES

1. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
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. Genetics*, **29**, 365–371.
3. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39(Suppl. 1)**, D1005–D1010.
4. Rayner,T.F., Rocca-Serra,P., Spellman,P.T., Causton,H.C., Farne,A., Holloway,E., Irizarry,R.A., Liu,J., Maier,D.S., Miller,M. *et al.* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.
5. 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.
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. Kodama,Y., Mashima,J., Kaminuma,E., Gojobori,T., Ogasawara,O., Takagi,T., Okubo,K. and Nakamura,Y. (2012) The DNA Data Bank of Japan launches a new resource, the DDBJ omics archive of functional genomics experiments. *Nucleic Acids Res.*, **40**, D38–D42.
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. Kapushesky,M., Adamusiak,T., Burdett,T., Culhane,A., Farne,A., Filippov,A., Holloway,E., Klebanov,A., Kryvych,N., Kurbatova,N. *et al.* (2012) Gene expression atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **40**, D1077–D1081.
10. 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.
11. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.
12. 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.
13. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
14. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
15. Goncalves,A., Tikhonov,A., Brazma,A. and Kapushesky,M. (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*, **27**, 867–869.
16. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
17. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
18. 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.
19. Shankar,R., Parkinson,H., Burdett,T., Hastings,E., Liu,J., Miller,M., Srinivasa,R., White,J., Brazma,A., Sherlock,G. *et al.* (2010) Annotare—a tool for annotating high-throughput biomedical investigations and resulting data. *Bioinformatics*, **26**, 2470–2471.
20. Gostev,M., Faulconbridge,A., Brandizi,M., Fernandez-Banet,J., Sarkans,U., Brazma,A. and Parkinson,H. (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.*, **40**, D64–D70.