D868–D872 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 10 November 2008*
doi:10.1093/nar/gkn889

ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression

Helen Parkinson[1], Misha Kapushesky[1], Nikolay Kolesnikov[1], Gabriella Rustici[1], Mohammad Shojatalab[1], Niran Abeygunawardena[1], Hugo Berube[2], Miroslaw Dylag[1], Ibrahim Emam[1], Anna Farne[1], Ele Holloway[1], Margus Lukk[1], James Malone[1], Roby Mani[1], Ekaterina Pilicheva[1], Tim F. Rayner[3], Faisal Rezwan[4], Anjan Sharma[1], Eleanor Williams[1], Xiangqun Zheng Bradley[1], Tomasz Adamusiak[1], Marco Brandizi[1], Tony Burdett[1], Richard Coulson[1], Maria Krestyaninova[1], Pavel Kurnosov[1], Eamonn Maguire[1], Sudeshna Guha Neogi[1], Philippe Rocca-Serra[1], Susanna-Assunta Sansone[1], Nataliya Sklyar[1], Mengyao Zhao[5], Ugis Sarkans[1] and Alvis Brazma[1,*]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [2]Institute for Information Technology, National Research Council Canada, Ottawa, Ontario, Canada, [3]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, [4]BioComputation Laboratory, University of Hertfordshire, Hatfield and [5]Wellcome Trust Sanger Institute, Hinxton, UK

Received September 12, 2008; Revised October 17, 2008; Accepted October 20, 2008

ABSTRACT

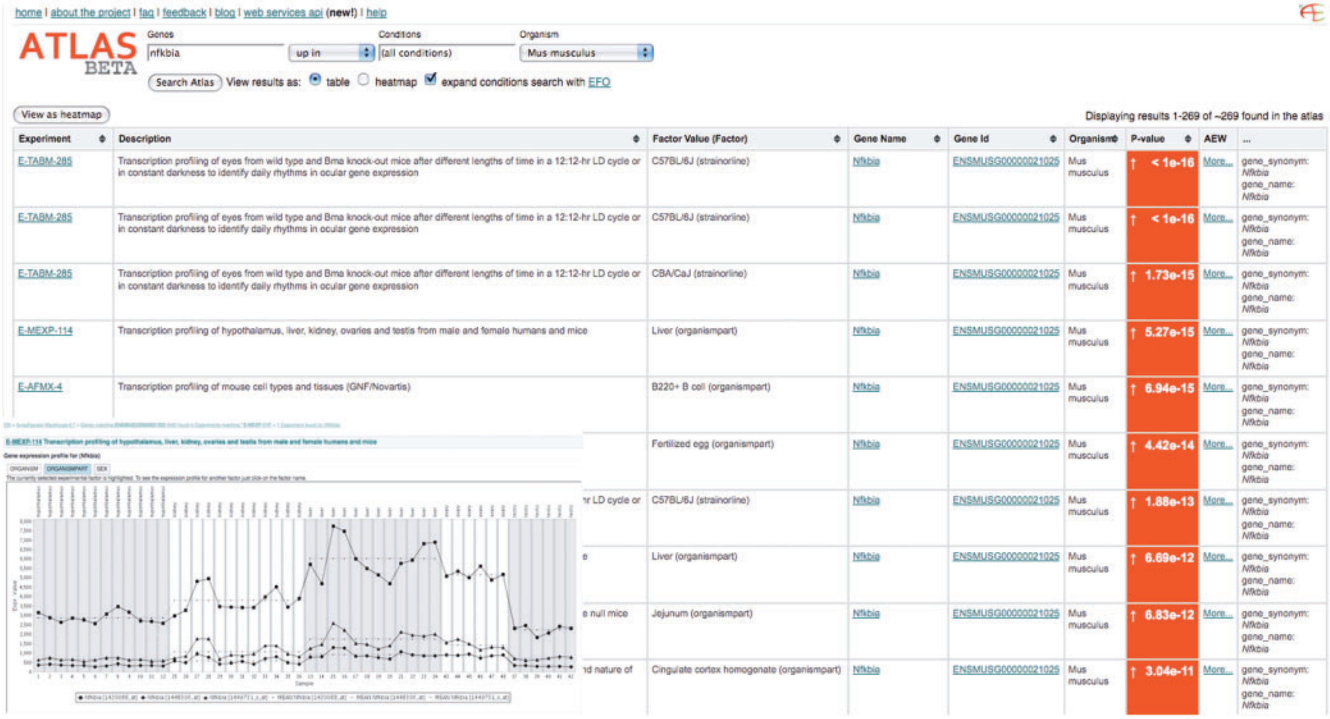
ArrayExpress http://www.ebi.ac.uk/arrayexpress consists of three components: the ArrayExpress Repository—a public archive of functional genomics experiments and supporting data, the ArrayExpress Warehouse—a database of gene expression profiles and other bio-measurements and the ArrayExpress Atlas—a new summary database and meta-analytical tool of ranked gene expression across multiple experiments and different biological conditions. The Repository contains data from over 6000 experiments comprising approximately 200 000 assays, and the database doubles in size every 15 months. The majority of the data are array based, but other data types are included, most recently—ultra high-throughput sequencing transcriptomics and epigenetic data. The Warehouse and Atlas allow users to query for differentially expressed genes by gene names and properties, experimental conditions and sample properties, or a combination of both. In this update, we describe the ArrayExpress developments over the last two years.

INTRODUCTION

The ArrayExpress Repository is one of the three recommended international repositories to archive publication-related functional genomics data (1). The ArrayExpress Repository of array-based data was launched in 2002 (2), and currently it contains data from over 6000 experiments (studies) and approximately 200 000 assays. The ArrayExpress Warehouse of gene expression profiles was added in 2005 (3,4), and currently it includes data from over 600 studies and 20 000 samples. There have been three major developments in the last two years. First, the new ArrayExpress Atlas of Gene Expression provides experimental condition-based queries and an overview of gene expression across multiple experiments (Figure 1). Second, we have established a procedure to import data from the Gene Expression Omnibus (GEO) (5) and over 3000 GEO data series (GSE) have been imported and released via the ArrayExpress interface. Third, ArrayExpress has started to accept data from ultra high-throughput sequencing (UHTS) experiments and the first data sets have been made public. Other improvements include more flexible data access, web services, the support of the MAGE-TAB format (6), direct import from

*To whom correspondence should be addressed. Tel: +44 1223 494658; Fax: +44 1223 494468; Email: brazma@ebi.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



home | about the project | faq | feedback | blog | web services api (new!) | help

ATLAS BETA

Genes: nfkbia — up in — Conditions: (all conditions) — Organism: Mus musculus

Search Atlas — View results as: ● table ○ heatmap ☑ expand conditions search with EFO

View as heatmap

Displaying results 1-269 of ~269 found in the atlas

Experiment	Description	Factor Value (Factor)	Gene Name	Gene Id	Organism	P-value	AEW	...
E-TABM-285	Transcription profiling of eyes from wild type and Bma knock-out mice after different lengths of time in a 12:12-hr LD cycle or in constant darkness to identify daily rhythms in ocular gene expression	C57BL/6J (strainorline)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ < 1e-16	More...	gene_synonym: Nfkbia gene_name: Nfkbia
E-TABM-285	Transcription profiling of eyes from wild type and Bma knock-out mice after different lengths of time in a 12:12-hr LD cycle or in constant darkness to identify daily rhythms in ocular gene expression	C57BL/6J (strainorline)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ < 1e-16	More...	gene_synonym: Nfkbia gene_name: Nfkbia
E-TABM-285	Transcription profiling of eyes from wild type and Bma knock-out mice after different lengths of time in a 12:12-hr LD cycle or in constant darkness to identify daily rhythms in ocular gene expression	CBA/CaJ (strainorline)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 1.73e-15	More...	gene_synonym: Nfkbia gene_name: Nfkbia
E-MEXP-114	Transcription profiling of hypothalamus, liver, kidney, ovaries and testis from male and female humans and mice	Liver (organismpart)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 5.27e-15	More...	gene_synonym: Nfkbia gene_name: Nfkbia
E-AFMX-4	Transcription profiling of mouse cell types and tissues (GNF/Novartis)	B220+ B cell (organismpart)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 6.94e-15	More...	gene_synonym: Nfkbia gene_name: Nfkbia
		Fertilized egg (organismpart)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 4.42e-14	More...	gene_synonym: Nfkbia gene_name: Nfkbia
	r LD cycle or	C57BL/6J (strainorline)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 1.88e-13	More...	gene_synonym: Nfkbia gene_name: Nfkbia
	e	Liver (organismpart)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 6.69e-12	More...	gene_synonym: Nfkbia gene_name: Nfkbia
	e null mice	Jejunum (organismpart)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 6.83e-12	More...	gene_synonym: Nfkbia gene_name: Nfkbia
	d nature of	Cingulate cortex homogenate (organismpart)	Nfkbia	ENSMUSG00000021025	Mus musculus	↑ 3.04e-11	More...	gene_synonym: Nfkbia gene_name: Nfkbia

E-MEXP-114 Transcription profiling of hypothalamus, liver, kidney, ovaries and testis from male and female humans and mice

Gene expression profile for (Nfkbia)

ORGANISM ORGANISMPART SEX

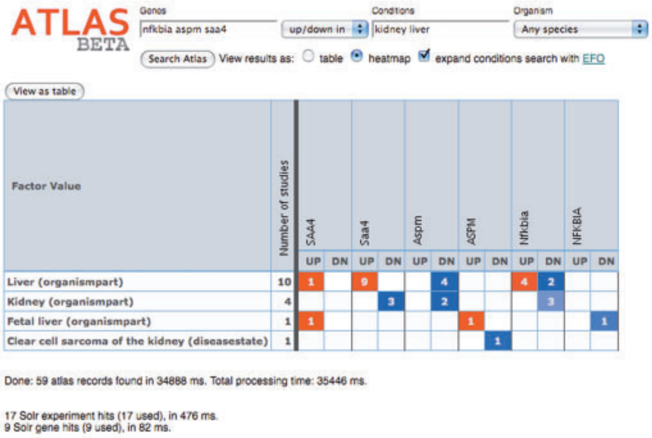
The currently selected experimental factor is highlighted. To see the expression profile for another factor just click on the factor name.

Figure 1. Table view from the ArrayExpress Atlas showing a query for up-regulation of the mouse gene Nfkbia in different experimental conditions. Columns from left to right: ArrayExpress experiment accession, experiment description, condition (factor value), gene name (linked to a gene view page), gene identifier, organism, *P*-value with an up/down indicator (red arrow for up, blue arrow for down) and a link to the ArrayExpress Warehouse for supporting data (inset bottom left).

ArrayExpress into Bioconductor http://bioconductor.org/packages/2.3/bioc/html/ArrayExpress.html, a simpler submission process for large experiments, and use of ontologies to annotate and query the data.

ARRAYEXPRESS ATLAS OF GENE EXPRESSION

The ArrayExpress Atlas of Gene Expression (http://www.ebi.ac.uk/microarray-as/atlas/) allows the user to query for condition-specific gene expression across multiple data sets. The user can query for a gene or a set of genes by name, synonym, Ensembl identifier, GO term or, alternatively, for a biological sample property or condition, e.g. tissue type, disease name, developmental stage, compound name or identifier. Queries for both genes and conditions are also possible. For example, the user can query for all 'DNA repair' genes up-regulated in 'cancer'. This returns a list of 'experiment, condition, gene' triplets each with a *P*-value and an up/down arrow characterizing the significance and direction of a gene's differential expression in a particular condition in an experiment. By default this list is ordered with most significant *P*-values on top. For example, a query for genes matching 'nfkbia', 'up' across all conditions in 'Mus musculus' returns a list of 230 matches in 30 different experiments and 100 different conditions where this gene is over-expressed (Figure 2). Each match is linked to detailed gene and experiment annotation in the Warehouse where expression profiles are presented as a graph of gene expression per sample and genes with similar expression profiles can be searched within an experiment. The raw experimental data files and complete experiment annotation are accessible through links to the ArrayExpress Repository.



ATLAS BETA

Genes: nfkbia aspm saa4 — up/down in — Conditions: kidney liver — Organism: Any species

Search Atlas — View results as: ○ table ● heatmap ☑ expand conditions search with EFO

View as table

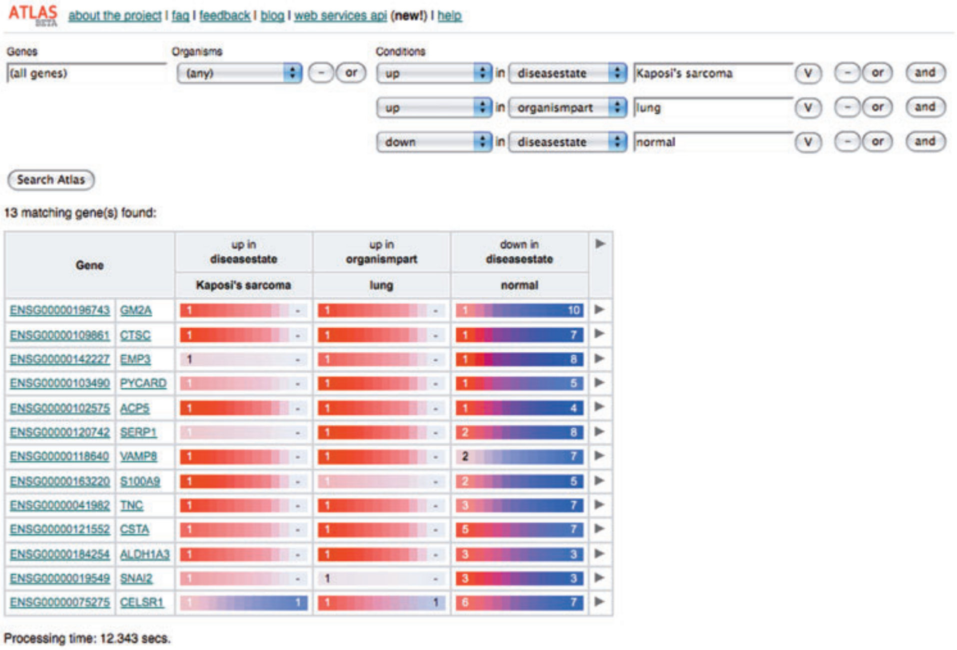
Factor Value	Number of studies	SAA4		Saa4		Aspm		ASPM		Nfkbia		NFKBIA	
		UP	DN	UP	DN	UP	DN	UP	DN	UP	DN	UP	DN
Liver (organismpart)	10	1		9			4			4	2		
Kidney (organismpart)	4				3		2				3		
Fetal liver (organismpart)	1	1						1					1
Clear cell sarcoma of the kidney (diseasestate)	1								1				

Done: 59 atlas records found in 34888 ms. Total processing time: 35446 ms.

17 Solr experiment hits (17 used), in 476 ms.
9 Solr gene hits (9 used), in 82 ms.

Figure 2. Heatmap view from the ArrayExpress Atlas showing an overview of expression of Nfkbia, Aspm and Saa4 genes in kidney and liver in all species.

As an alternative to presenting a ranked list of triplets, the Atlas heat map view provides a summarized table of gene expression patterns matching the query. Genes are listed in columns, conditions in rows, and the table cells are coloured red for over-expression, blue for under-expression and the number of independent studies where the pattern is observed is displayed within each


ATLAS BETA
about the project | faq | feedback | blog | web services api (new!) | help
Search Atlas
V
−
or
and

Genes: (all genes)
Organisms: (any)
Conditions:
up in diseasestate Kaposi's sarcoma
up in organismpart lung
down in diseasestate normal

13 matching gene(s) found:

Gene		up in diseasestate	up in organismpart	down in diseasestate	▶
		Kaposi's sarcoma	lung	normal	
ENSG00000196743	GM2A	1 -	1 -	1 10	▶
ENSG00000109861	CTSC	1 -	1 -	1 7	▶
ENSG00000142227	EMP3	1 -	1 -	1 8	▶
ENSG00000103490	PYCARD	1 -	1 -	1 5	▶
ENSG00000102575	ACP5	1 -	1 -	1 4	▶
ENSG00000120742	SERP1	1 -	1 -	2 8	▶
ENSG00000118640	VAMP8	1 -	1 -	2 7	▶
ENSG00000163220	S100A9	1 -	1 -	2 5	▶
ENSG00000041982	TNC	1 -	1 -	3 7	▶
ENSG00000121552	CSTA	1 -	1 -	5 7	▶
ENSG00000184254	ALDH1A3	1 -	1 -	3 3	▶
ENSG00000019549	SNAI2	1 -	1 -	3 3	▶
ENSG00000075275	CELSR1	1 1	1 1	6 7	▶

Processing time: 12.343 secs.

Figure 3. ArrayExpress Atlas advanced query interface showing all genes up-regulated in Kaposi's sarcoma diseased lung tissue and down-regulated in normal samples.

cell (Figure 3). More complex queries combining several conditions are available via the advanced Atlas interface.

The Atlas data content is produced using a meta-analytical approach which uses the Bioconductor module limma (7). It is applied to each experiment in the ArrayExpress Warehouse to calculate moderated t statistics that simultaneously test the significance of differential expression in each condition versus the mean level of expression across all conditions for each gene. We then adjust the computed P-values to control the false discovery rate, using the Benjamini–Hochberg method (8), and aggregate the computed multiple test statistics for interpretation. The high quality of annotation and curation of experiments in the ArrayExpress Warehouse allows the application of this generic method to data from multiple platforms. Although the majority of the Atlas data at present is on the Affymetrix platform, the method is platform independent, and the Atlas also includes data from two-channel arrays. The ArrayExpress Warehouse and Atlas have a monthly data release with data derived from the ArrayExpress Repository content. Each release includes new experiments and a complete re-annotation of the genes present based on Ensembl and Uniprot, together with updated gene rankings.

DATA IMPORT FROM GEO

ArrayExpress now integrates gene expression data produced on Affymetrix and Agilent array platforms from GEO. This allows users to view and search GEO and ArrayExpress data from a common interface and access the data in the standard MAGE-TAB format. GEO data sets (GDS), using Affymetrix and Agilent catalogue arrays, are imported and converted from GEO SOFT to MAGE-TAB format. The data files are then checked and free text information is text mined using Whatizit (9) and a custom ArrayExpress dictionary derived from the NCI Thesaurus (10) with local extensions for non-human terms. Imported and text-mined datasets are then curated and the annotations mapped to ontology terms. GSE on Affymetrix and Agilent catalogue arrays are also imported and curated when GEO releases these as GDS or on request by a user community. The original GEO format accession numbers are retained as secondary accessions and links to the original data in GEO provided from the ArrayExpress interface. To date, there are 3700 GEO-derived experiments comprising more than 100 000 hybridizations present in the ArrayExpress repository. Code for parsing GEO SOFT format to MAGE-TAB, text-mining utilities and the custom dictionary are available (http://sourceforge.net/project/showfiles.php?group_id = 120325). Mappings from SOFT to MAGE-TAB formats have been validated by GEO (T. Barrett, personal communication).

Once GDS are present in the ArrayExpress Repository, we apply Bioconductor quality metrics and perform scoring for MIAME compliance (11) to decide which data sets are appropriate for integration into the ArrayExpress Data Warehouse/Atlas. Curated GEO data that satisfy both the quality metrics and MIAME compliance and have curated experimental factors are included in the ArrayExpress Data Warehouse and consequently made available via Atlas queries.

NEW HIGH-THROUGHPUT SEQUENCING AND OTHER SUPPORTED DATA TYPES

ArrayExpress accepts data generated on all array-based technologies, including gene expression, protein array, ChIP-chip and genotyping. More recently, data from transcriptomic and related applications of UHTS technologies such as Illumina (SOLEXA Ltd, Saffron Walden, UK) (12), and 454 Life Sciences (Roche, Branford, Connecticut) are also accepted. For Solexa data FASTQ files, sample annotation and processed data files corresponding to transcription values per genomic location are submitted and curated to the emerging standard MINSEQE (http://www.mged.org/minseqe) and instrument-level data are stored in the European Short Read Archive (http://www.ebi.ac.uk/embl/Documentation/ENA-Reads.html). The first sequencing-based data sets to be made public are the transcriptomics landscape for *Schizosaccharomyces pombe* and epigenomics analysis of human (accession numbers E-MTAB-5 and E-TABM-482) (13,14). Spreadsheet templates are available for submission of these data from the ArrayExpress homepage.

The ArrayExpress Warehouse now includes gene expression profiles from *in situ* gene expression measurements, as well as other molecular measurement data from metabolomics and protein profiling technologies. Where *in situ* and array-based gene expression data are available for the same gene, these are displayed in the same view and links to the multispecies 4DXpress database of *in situ* gene expression (15) are provided.

OTHER DEVELOPMENTS

In addition to integrating processed data from multiple array platforms in the ArrayExpress Atlas, we have also performed a per platform integration using a re-annotation, data quality assessment and re-normalization approach. A large data set of more than 5000 hybridizations and 370 different biological conditions on the Affymetrix U133A platform is now available. The meta-analyses indicate that despite these data originating from multiple laboratories, the biological signal in these data is significantly stronger than the laboratory effects and new biological insights can be obtained from this approach (Lukk *et al.*, manuscript in preparation). All raw and normalized data are available for this dataset (accession number E-TABM-185). Similar data sets have been produced for the mouse Affymetrix platforms U74Av2, MOE430A and 430 2.0 (accession numbers E-MTAB-26, E-MTAB-27, E-MTAB-28).

All ArrayExpress data are now available for download in MAGE-TAB format. To aid bioinformaticians and other users interested in large-scale functional genomics analysis, a Bioconductor package called ArrayExpress (http://www.bioconductor.org/packages/2.3/bioc/html/ArrayExpress.html) has been developed in collaboration with the Huber Group (EBI). This package allows the direct import of MAGE-TAB files into Bioconductor as native ExpressionSet objects, compatible with existing Bioconductor data analysis and visualization modules for this environment.

The ArrayExpress Atlas, Repository and Warehouse have web service APIs, enabling programmatic queries. ArrayExpress can also be queried along with all EBI core databases via the EBI general query interface ‘EB-eye’. The ArrayExpress submission tools, MIAMExpress and Tab2MAGE, are undergoing continuous improvement to facilitate submissions of large experiments, to work with MAGE-TAB files, and to accept UHTS-based transcriptomics data.

To improve queries of the ArrayExpress Atlas, we have developed an application ontology called the Experimental Factor Ontology (EFO). EFO version 0.6 (16) currently contains 1078 terms in an ‘is-a’ and ‘part-of’ hierarchy including diseases, multi-species anatomy, compounds and cell-type terms. It maps to several non-orthogonal ontologies, such as those for human anatomy, the Disease Ontology (17), the Cell Type Ontology (18) and the NCI Thesaurus (10). Use of the EFO allows tuning of the ontology based on analysis of user queries and provision of annotation at an appropriate level of granularity for the database content. The EFO is deployed in the Atlas interface where queries can be expanded via the hierarchies. For example, a query for the condition ‘cancer’ will also retrieve conditions ‘sarcoma’, ‘carcinoma’ and other cancers. The EFO is available from the EBI Ontology Lookup Server-OLS (19) and is available in OBO and OWL formats (http://www.ebi.ac.uk/microarray-srv/efo/).

FUTURE

There are two main challenges for ArrayExpress in the next two years—first to increase the volume of data in the Warehouse and Atlas, and second to build a standard pipeline for accepting and presenting UHTS-based data via the ArrayExpress Atlas. The underlying ArrayExpress architecture is undergoing a major re-development, which will provide improved performance to deal with large data sets and integration of multi-technology studies.

Currently, the ArrayExpress Warehouse/Atlas contains 10% of the data that are stored in the Repository and increasing this proportion in the next 12 months is a priority. An infrastructure is being developed that will allow the ArrayExpress curators to assess the appropriateness of new submissions for the Atlas at the point of the submission to the Repository, and to curate and load these data into a private Atlas instance. These data sets will then be automatically released in the Atlas as they become public. UHTS-based functional genomics data are already accepted and made available via the Repository. In the future, we will develop data analysis pipelines for these data sets for integration into the Atlas.

A new sample and experimental metadata database called the Bio Investigation Index (http://www.ebi.ac.uk/net-project/projects.html) is under development and will create a common structured representation and storage mechanism for metadata across several databases at the EBI. The new infrastructure also includes an emerging format for multi platform-based studies, ISA-TAB (20). Tools for creating ISA-TAB are under development, as well as converters between ISA-TAB and MAGE-TAB.

The ArrayExpress Atlas will undergo substantial user interface improvements, which will include a seamless integration of the Atlas with our online data analysis tool Expression Profiler (21). The code and ranking methodology will be made available via the Atlas documentation and a Bioconductor package. These will be the subject of a future publication. Finally, although the Repository, Warehouse and Atlas are distinct databases with different missions, a common user interface is under development and will be available in 2009.

ACKNOWLEDGEMENTS

We would like to thank Wolfgang Huber and Audrey Kauffman for their work on the ArrayExpress and arrayQualityMetrics Bioconductor packages, Torsten Heinrich for assistance with import of *in situ* gene expression data and Tanya Barrett of GEO for assistance with SOFT to MAGE-TAB mapping. We would also like to thank the ArrayExpress Scientific Advisory Board.

FUNDING

The European Molecular Biology Laboratory; the European Commission grants FELICS (021902); EMERALD (LSHG-CT-2006-037689); DIAMONDS (LSHG-CT-2004-512143); BioSapiens (LSHG-CT-2003-503265); MolPAGE (LSHG-CT-2004-512066); MUGEN (LSHG-CT-2005-005203); Carcinogenomics (LSHB-2006-037712); Nutrigenomics (NoE-503630); Biotechnology and Biological Sciences Research Council (BB/E025080/1); the MAGE grant from National Human Genome Research Institute and National Institute of Biomedical Imaging and Bioengineering from the National Institute of Health (1 P41 HG003619-01). Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. 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**, 1276–1277.
2. 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.
3. 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.
4. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.* (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
5. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
6. Rayner,T., Rocca-Serra,P., Spellman,PT, 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.
7. Gentleman,R., Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.
8. Benjamini,Y. and Hochberg,Y. (1995) Controlling for the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
9. Rebholz-Schuhmann,D., Arregui,M., Gaudan,S., Kirsch,H. and Jimeno,A. (2008) Text processing through web services: calling Whatizit. *Bioinformatics*, **24**, 296–298.
10. Fragoso,G., de Coronado,S., Haber,M., Hartel,F. and Wright,L. (2004) Overview and utilization of the NCI Thesaurus. *Comp. Funct. Genomics*, **5**, 648–654.
11. Brazma,A. and Parkinson,H. (2006) ArrayExpress service for reviewers/editors of DNA microarray papers. *Nat. Biotechnol.*, **24**, 1321–1322.
12. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
13. Wilhelm,B., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
14. Down,T.A., Rakyan,V.K., Turner,D.J., Flicek,P., Li,H., Kulesha,E., Gräf,S., Johnson,N., Herrero,J., Tomazou,E.M. *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.*, **26**, 779–785.
15. Haudry,Y., Berube,H., Letunic,I., Weeber,P.D., Gagneur,J., Girardot,C., Kapushesky,M., Arendt,D., Bork,P., Brazma,A. *et al.* (2008) 4DXpress: a database for cross-species expression pattern comparisons. *Nucleic Acids Res.*, **36**, D847–D853.
16. Malone,J., Zheng-Bradley,H.J., Rayner,T. and Parkinson,H. (2008) Developing an application focused experimental factor ontology: embracing the OBO Community. In *Proceedings of the Eleventh Annual Bio-ontologies Meeting*. Toronto, San Diego, USA, pp. 21–24.
17. Dyck,P. and Chisholm,R. (2003) Disease ontology: structuring medical billing codes for medical record mining and disease gene association. In *Proceedings of the Sixth Annual Bio-ontologies Meeting*. Brisbane, San Diego, USA, pp. 53–55.
18. Bard,J., Rhee,S.Y. and Ashburner,M. (2005) An ontology for cell types. *Genome Biol.*, **6**, R21.
19. Cote,R.G., Jones,P., Apweiler,R. and Hermjakob,H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.
20. Sansone,S.A., Rocca-Serra,P., Brandizi,M., Brazma,A., Field,D., Fostel,J., Garrow,A.G., Gilbert,J., Goodsaid,F., Hardy,N. *et al.* (2008) The first RSBI (ISA-TAB) workshop: "Can a simple format work for complex studies?". *OMICS*, **12**, 143–149.
21. Kapushesky,M., Kemmeren,P., Culhane,A.C., Durinck,S., Ihmels,J., Körner,C., Kull,M., Torrente,A., Sarkans,U., Vilo,J. *et al.* (2004) Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucleic Acids Res.*, **32**, W465–W470.