

Gabriella Rustici (v3, 31/03/2009)

ArrayExpress for Beginners

The ArrayExpress (AE) database is a public curated repository of **transcriptomics** data. As of March 2009 the database holds approximately 230,000 assays, from over 7700 separate studies (or experiments) related to over 200 different species.

You will learn about:

- The basics of ArrayExpress – what it is and when to use it?
- How to query the database
- How to interpret the data
- How to export data from the database to other applications

Contents:

- 1 What is ArrayExpress and when to use it?
- 2 How to query the ArrayExpress Warehouse
- 3 How to query the ArrayExpress Archive
- 4 How to query the ArrayExpress Advance Interface
- 5 How to query the ArrayExpress Archive by experiment accession number

1 What is ArrayExpress and when to use it?

ArrayExpress is a public repository for transcriptomics data, which is aimed at storing **MIAME**- and **MINSEQE**- compliant data in accordance with the Microarray and Gene Expression Data (MGED) Society recommendations (<http://www.mged.org/>) [1].

AE resource consists of two databases: (1) AE Archive, which stores well annotated microarray data typically supporting journal publications, and (2) AE Warehouse of gene expression profiles, which contains additionally curated subsets of data from the AE Archive and enables the user to query **gene expression profiles** by gene names, properties and profile similarity [2].

You can go straight to the EBI's ArrayExpress page by using the ArrayExpress link from the EBI front page: <http://www.ebi.ac.uk> (Fig. 1).



This work is licensed under the Creative Commons Attribution-Share Alike 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/3.0/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

The screenshot shows the EMBL-EBI website interface. At the top, there is a search bar with 'All Databases' selected and a 'Go' button. Below the search bar is a navigation menu with links for Databases, Tools, EBI Groups, Training, Industry, About Us, and Help. The main content area is divided into several sections:

- Data Resources & Tools:** A grid of links including EMBL-BANK, UniProt, ArrayExpress (circled in red), Ensembl, InterPro, PDB-EBI, Genomes, Nucleotide Sequences, Protein Sequences, Macromolecular Structures, Small Molecules, Gene Expression, Molecular Interactions, Reactions & Pathways, Protein Families, Enzymes, Literature, Taxonomy, Ontologies, Sequence Similarity & Analysis, Pattern & Motif Searches, Structure Analysis, Text Mining, and Downloads.
- European Bioinformatics Institute:** A central banner with a building image.
- About the EBI:** Links for Research, PhD Studies, Training, Industry Support, Group & Team Leaders, EBI Funders, User Support, EBI Mission, People, Events at the EBI, and How to Find us.
- EBI hosted EU Project Websites:** Links for BioSapiens, E-MeP, ELIXIR, EMBRACE, EMERALD, ENFIN, FELICS, and SYMBIOmatics.
- Hands-on Courses:** Information about registration for proteomics resources and interactions/pathways courses.
- Research Highlights:** A news item from June 19, 2008, about a new computational tool for evolution.
- Latest News:** A news item from May 28, 2008, about funding for the ELIXIR project.

Fig. 1: Link to AE from the EBI front page (<http://www.ebi.ac.uk/>)

2 How to query the ArrayExpress Warehouse

The use of the AE Warehouse is straight forward – enter the name, ID or a property of a gene or several genes, retrieve the list of experiments where the given gene has been studied, and zoom into its expression profile.

To familiarise you with the AE query form, we will perform a simple search, querying for the expression profiles of a single gene.

1. Open the AE homepage, at <http://www.ebi.ac.uk/arrayexpress/>, in a Web browser
2. In the 'Expression Profiles' box, on the right-hand side of the page, type in any gene name, e.g. **nfkbia**, and 'leukemia' as keyword (as shown in Fig. 2)
3. Select species, e.g. *Homo sapiens*, in the 'Species' dropdown menu and click the 'query' button.

Fig. 2: The ArrayExpress query window (<http://www.ebi.ac.uk/arrayexpress/>)

The interface returns the list of all experiments (studies) in the AE Warehouse where the selected gene has been studied (Fig. 3). Experiments are ordered by *p-value*, in ascending order. The *p-value* is based on the correlation between **experimental factors** values and gene expression values and it is calculated using several methods, including a linear model in the **Bioconductor** package LIMMA [3]. For each **experiment**, a short description, a list of **experimental factors** and the experimental set up (type) are provided. In addition, a thumbnail image shows the behaviour of the selected gene in each **experiment** retrieved. At a glance the user can now decide which **experiment** might be interesting to further viewing.

Fig. 3: Output window after querying the AE Warehouse for the expression profiles of a particular gene (e.g. *nfkbia*)

Click on the thumbnail image of the expression profile in one of the **experiments**, e.g. E-AFMX-5. In the graph that now shows (as displayed in Fig. 4), the X axis represents all samples in this study, grouped by **experimental factor** and the Y axis the expression levels for *nfkbia* in each **sample**. Explore the dependency of the expression levels on different **experimental factors**. **Experimental factors** are the main experimental variables studied in a particular study. For instance, **experiment** E-AFMX-5 has three **experimental factors** – cell type, disease state and organism part. Select the **experimental factor** ‘cell type’ and observe that, under this condition, *nfkbia* has notably higher expression values for the cell type CD33+ myeloid, than for instance CD4+ T cells (Fig. 4). The black line represents the expression value for the **Affymetrix probe** 201502_s_at, for *nfkbia*. The dotted lines represent the mean expression values.

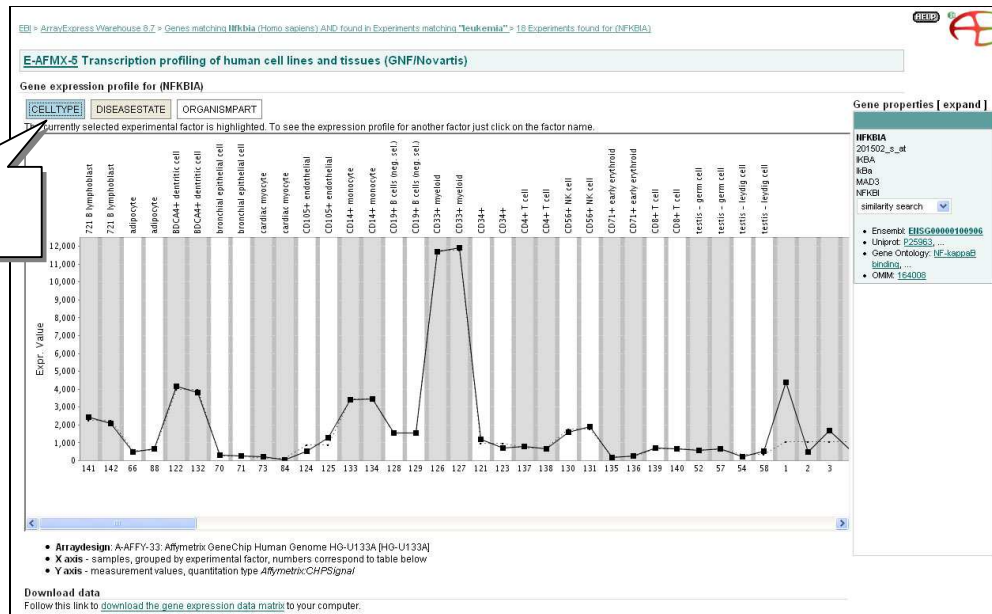


Fig. 4: Zoomed-in view of a particular experiment. The main graph shows the expression profile of the selected gene (e.g. *nfkbia*), for all experimental samples, based on the selected experimental factor

Scroll down the page for more information about the **sample** properties. In the table provided, the **sample** number in the first column corresponds to the **sample** number on the X axis of the graph (Fig. 5). Please note that the expression values are measured in abstract units as supplied by the submitter. For instance, E-AFMX-5 uses **Affymetrix** platform and **MAS5 normalisation method**. For more information about the particular **normalisation** protocols used in each individual **experiment**, click on the **experiment** accession number. This will open the link to the respective dataset entry in the AE Archive, which contains all the information related to the selected study.

Sample	Species	Age	Clinical history	Developmental stage	Disease state	Observation	Organism part	Sex	Target cell type	Cell type [ef]
1	Homo sapiens	30-40 years since birth	sudden death	adult		Caucasian	lung	mixed_sex		
2	Homo sapiens	25 years since birth	sudden death	adult		Asian	heart	male		
3	Homo sapiens	21-50 years since birth	trauma	adult		Caucasian	prostate	male		
4	Homo sapiens			adult			uterus			
5	Homo sapiens	27 years since birth	sudden death	adult		Asian	liver	male		
6	Homo sapiens			adult	promyelocytic leukemia HL-60	Caucasian	tumor			
7	Homo sapiens				chronic myelogenous leukemia K562		tumor			
8	Homo sapiens			adult			spinal cord			
9	Homo sapiens	20-33 weeks since fecundation	spontaneously aborted	fetus		Caucasian	fetal brain	mixed_sex		
10	Homo sapiens	20-61 years since birth	sudden death	adult		Caucasian	lymph node	mixed_sex		
11	Homo sapiens	19 years since birth		adult	lymphoblastic leukemia MOLT-4	Caucasian	tumor	male		
12	Homo sapiens			adult	promyelocytic leukemia HL-60	Caucasian	tumor			
13	Homo sapiens				chronic myelogenous leukemia K562		tumor			

Fig. 5: Table of sample properties for the selected experiment

On the top right hand side of the same page (as shown in Fig. 4), click on the **similarity search**, and from the drop-down menu select the 'find 3 closest genes' option. This will select the 3 most

similarly expressed genes and add their expression profiles to the selection, next to the **nfkbia** profiles (Fig. 6). You will find that the expression patterns for genes IER2, FOS and JUN closely resemble the behavior of **nfkbia**. Click on 'expand' Gene Properties, and follow the links from the expanded view to retrieve additional information about these genes, in ENSEMBL, Uniprot, **QuickGO**, **OMIM** and **4DXpress** databases. Finally, by clicking on the 'download the gene expression data matrix' link located below the graph (as circled in Figure 6), one can obtain the numerical expression values of the selected genes for further analysis.

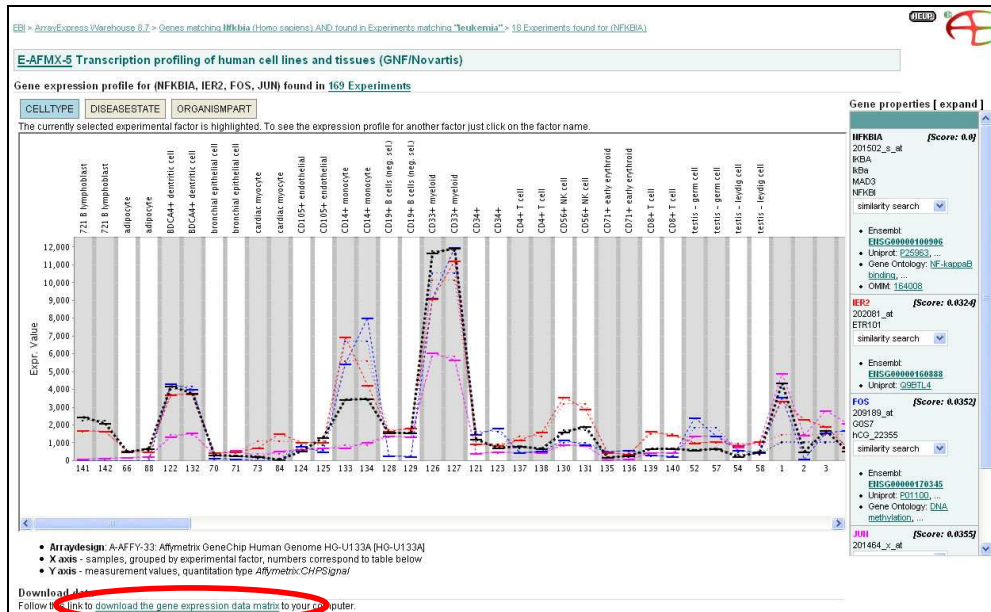


Fig. 6: Similarity search output window. The expression profile of the selected gene (e.g. **nfkbia**) is plotted together with the ones of the 3 genes showing the closest similarity in expression pattern, within the same experiment. The corresponding gene symbols are listed on the right (**Ier2**, **Fos**, and **Jun**)

It is also possible to query for more than one gene at the time. For example, enter two or more comma separated gene names (e.g. '**Ephb3**, **Nfkbia**'), select species *Mus musculus* and click the 'query' button (as shown in fig. 7).

Fig. 7: The ArrayExpress query window (<http://www.ebi.ac.uk/arrayexpress/>)

If more than one gene is selected, the query tries to match the gene names exactly (Fig. 8). The user will be prompted to an intermediate window where a list of matching genes found is provided, together with a list of matching **experiments**. This is particularly useful when there is more than one gene matching your original query. Toggle the genes of interest (in this case both of them) and then click display. The familiar **experiment** thumbnails page will now be displayed and can be browsed as previously described.

Display Expression Profiles

Select one or more genes and experiments to visualize expression Display >

Or click on 'show gene exp. profiles' to visualize expression of the desired gene

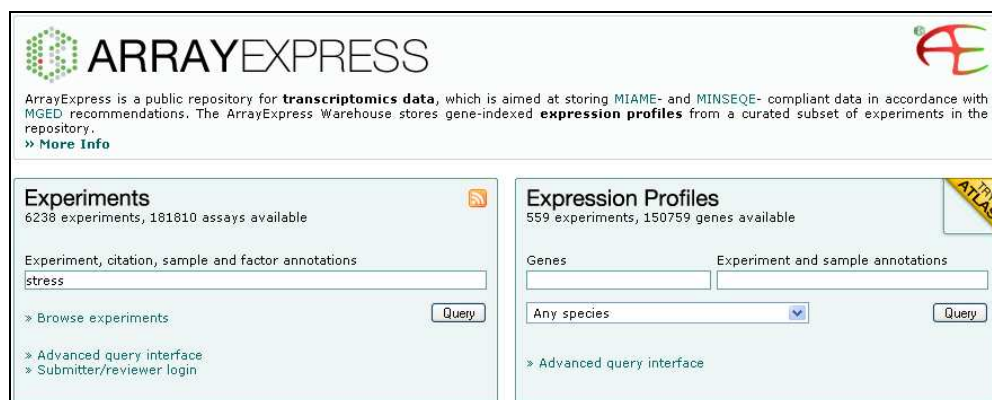
Matching Genes				Matching Experiments			
Displaying genes 1 - 2 of 2 found				Displaying 36 Experiments			
Toggle all	Gene	Ensembl	UniProt	RefSeq	Toggle all	Experiment	Title
1. <input checked="" type="checkbox"/>	Itih3 (Mus musculus) <ul style="list-style-type: none"> Gene Ontology: cytosol, negative regulation of Notch signaling pathway, negative regulation of myeloid cell differentiation, protein binding, regulation of cell proliferation, protein import into nucleus, translocation, nucleus Synonyms: Ikba, Nfkbia InterPro Terms: Ankyrin Keywords: ANK repeat, Phosphorylation, Repeat, Ub1 conjugation Experiment: E-AFMX-1, E-AFMX-4, E-MEXP-1005, E-MEXP-1030, E-MEXP-1040, E-MEXP-114, E-MEXP-255, E-MEXP-420, E-MEXP-438, E-MEXP-453, E-MEXP-454, E-MEXP-459, E-MEXP-490, E-MEXP-558, E-MEXP-565, E-MEXP-634, E-MEXP-700, E-MEXP-701, E-MEXP-710, E-MEXP-731, E-MEXP-748, E-MEXP-774, E-MEXP-82, E-MEXP-842, E-MEXP-878, E-MEXP-886, E-MEXP-891, E-MEXP-892, E-MEXP-893, E-MEXP-917, E-MEXP-923, E-MEXP-939, E-TABM-102, E-TABM-163, E-TABM-199, E-TABM-229 	EHSMSG00000021025	Q3U9W8 Q3UB40 Q9Z1E3	NM_010907	1. <input checked="" type="checkbox"/>	E-MEXP-453	Transcription profiling of mouse motoneurons during progression of the disease in transgenic SOD1 (G93A) mice that develop motoneuron loss
					2. <input checked="" type="checkbox"/>	E-MEXP-893	Transcription profiling of hepatocytes from mice that are hyperglycaemic, obese and insulin resistant as a result of being fed a high fat diet
					3. <input checked="" type="checkbox"/>	E-AFMX-1	Transcription profiling of human, chimp and mouse brain
					4. <input checked="" type="checkbox"/>	E-MEXP-774	Transcription profiling of mouse TS-L1 preadipocytes treated with the steroid hormone dexamethasone in biological and technical replicates to illustrate validation methods
					5. <input checked="" type="checkbox"/>	E-MEXP-82	Transcription profiling time course toxicogenomic profiles in CD-1 mice livers after non-toxic and toxic paracetamol administration
2. <input checked="" type="checkbox"/>	Ephb3 (Mus musculus) <ul style="list-style-type: none"> Gene Ontology: axon guidance, axon guidance receptor activity, extracellular space, protein serine/threonine kinase activity, transmembrane-ephrin receptor activity, integral to plasma membrane Synonyms: Ephb3, Etk2, Mtk5, Sek4 InterPro Terms: Ephrin receptor, ligand binding, Fibronectin, type III, Fibronectin, type III subdomain, Galactose-binding like, Protein kinase, Receptor tyrosine kinase, class V, Sterile alpha motif-type, Tyrosine protein kinase, active site, Tyrosine protein kinase, Sterile alpha motif SAM, Protein kinase like, Growth factor receptor, Fibronectin type 	EHSMSG00000005958	P54754 Q3T277 Q60669 Q91YS9	NM_010143			

Fig. 8: Gene selection page. When more than one gene matches the query, this window allows refining the search, querying for multiple genes or restricting the search to perfect matches only

3 How to query the ArrayExpress Archive

The AE Archive allows the user to browse or query the **experiments** via free text search (e.g. **experiment** accession numbers, authors, laboratory, publication, key words), and filter the experiments retrieved by species or array design or **experiment** type. Once the desired **experiment** is identified, the user can find more information about the samples, protocols used, experimental design, etc. and most importantly can export the data associated with the selected **experiment**.

1. Go to the AE main homepage, at <http://www.ebi.ac.uk/arrayexpress/>
2. In the 'Experiments' box, on the left-hand side of the page, type in a word or a phrase or **GO** term by which you want to retrieve the **experiments**, e.g. 'stress' and click the 'Query' button (as shown in Fig. 9)



ARRAYEXPRESS

ArrayExpress is a public repository for **transcriptomics data**, which is aimed at storing MIAME- and MINSEQE- compliant data in accordance with MGED recommendations. The ArrayExpress Warehouse stores gene-indexed **expression profiles** from a curated subset of experiments in the repository.
» [More Info](#)

Experiments
6238 experiments, 181810 assays available

Experiment, citation, sample and factor annotations
stress

» [Browse experiments](#)

» [Advanced query interface](#)
» [Submitter/reviewer login](#)

Expression Profiles
559 experiments, 150759 genes available

Genes Experiment and sample annotations

Any species

» [Advanced query interface](#)

Fig. 9: The ArrayExpress query window (<http://www.ebi.ac.uk/arrayexpress/>)

This will bring up a window with a list of **experiments** in the reverse order of their publication dates in the AE Archive (Fig. 10). One can increase the number of **experiments** per page, by changing the default in the top right corner up to 500 per page.

For each **experiment** the following information are displayed:

- an **experiment** accession number (ID). This is a unique identifier assigned to each **experiment** by the AE curation staff. The accession number can also be used to query the Archive;
- a title, with a brief description of the **experiment**;
- the number of assays associated with the **experiment**;
- data availability, as processed or raw data, and links to the **ArrayExpress Atlas of Gene Expression**, when available.

AE unique experiment ID

Curated title of experiment

Number of assays in each experiment

Species investigated (can be multiple)

The date when the data were loaded in the Archive

Direct link to the AE Atlas of Gene Expression.

The ✓ icon indicates that the selected experiment is available in the AE Atlas.

The total number of experiments retrieved

The list of experiments retrieved can be printed, saved as Tab-delimited format or exported to Excel or as RSS feed

The total number of assays retrieved

The direct link to the processed data as a .zip file. An icon indicates that this type of data is available.

The direct link to the raw data. An icon indicates that this type of data is available.

A wedge shaped icon indicates Affymetrix arrays were used.

749 experiments, 17840 assays. Displaying experiments 1 to 25. Pages: 1 2 3 4 5 6 7 8 9 10

Fig. 10: Output window after querying the AE Archive for a particular set of experiments, using a word or phrase (e.g. stress). The total number of experiments and corresponding samples retrieved appears at the bottom of the page.

It is possible to apply additional filtering to select a particular species and/or array platform and/or experiment type by using the drop down menus at the top of the page (Fig. 10). In this case, filter on the species *Schizosaccharomyces pombe* and click the 'Query' button. The updated window will now show only the **experiments** present in the database which are related to 'stress' for the selected species, *Schizosaccharomyces pombe* (Fig. 11).

Experiment, citation, sample and factor annotations [clear] Filter on [reset] Display options [reset]

stress Schizosaccharomyces pombe 25 experiments per page

Match whole words Loaded in ArrayExpress Atlas Any array Detailed view

Submitter/reviewer login ArrayExpress Browser Help Query

ID	Title	Assays	Species	Date	Processed	Raw	Atlas
E-MTAB-5	High throughput sequencing of fission yeast to survey the dynamic repertoire of a eukaryotic transcriptome at	26	Schizosaccharomyces pombe	2008-12-03			
E-MTAB-18	Transcription profiling of fission yeast under multiple conditions, including exponential proliferation, meiotic diffe	25	Schizosaccharomyces pombe	2008-04-15			
E-TABM-447	Transcription profiling of wild type and Prct1 or Atf1 knock out fission yeast in the presence and absence of oxid	30	Schizosaccharomyces pombe	2008-03-14			
E-MEXP-1083	Transcription profiling of wild type and mutant fission yeast exposed to different intensities of H2O2 and two oth	153	Schizosaccharomyces pombe	2007-12-05			
E-TABM-298	Transcription profiling time series of meidelta mutant fission yeast during meiosis	12	Schizosaccharomyces pombe	2007-08-07			
E-TABM-299	Transcription profiling of fission yeasts over-expressing transcription factors controlling sexual differentiation	8	Schizosaccharomyces pombe	2007-08-07			
E-TABM-300	Transcription profiling time series of repidelta mutant fission yeast during meiosis	5	Schizosaccharomyces pombe	2007-08-07			
E-TABM-301	Transcription profiling of wild type, rsv1 deletion and SPBC11-5.14 deletion mutant fission yeast	4	Schizosaccharomyces pombe	2007-08-07			
E-MEXP-1127	Transcription profiling of three strains of fission yeast treated with hydroxyurea for different lengths of time	24	Schizosaccharomyces pombe	2007-06-22			
E-TABM-120	Transcription profiling of fission yeast in response to changes in copper and iron levels	31	Schizosaccharomyces pombe	2007-03-01			
E-MEXP-29	Transcription profiling of stress treated fission yeast (Stressors: compound: heavy metal, oxidation, alkylation a	67	Schizosaccharomyces pombe	2004-10-07			
E-MEXP-176	Transcription profiling of fission yeast treated with hydrogen peroxide in fission yeast wild type and csx1 mutant	12	Schizosaccharomyces pombe	2004-10-01			
E-SNGR-3	Transcription profiling of fission yeast meiosis and sporulation	14	Schizosaccharomyces pombe				
E-SNGR-7	Transcription profiling of fission yeast meiosis and sporulation	25	Schizosaccharomyces pombe				

14 experiments, 436 assays.

Fig. 11: Output window after querying the AE Archive for a particular set of experiments, using a word or phrase (e.g. stress) and selecting a species (e.g. Schizosaccharomyces pombe)

It is possible to expand the **experiment** view by clicking on the plus sign next to the experiment ID. For instance, expand the **experiment** view for E-MEXP-29. Additional information is provided in the new window together with extremely useful links to **experiment** annotation and data retrieval (Fig. 12).

The screenshot displays the EBI ArrayExpress interface for experiment E-MEXP-29. The main content area is divided into several sections, each highlighted by a callout box:

- Experiment description:** Points to the title "Transcription profiling of stress treated fission yeast" and the detailed description of the experiment.
- MIAME score:** Points to the MIAME score section, which shows a score of 5 and a list of criteria that are fulfilled (Array designs, Protocols, Factors, Processed data, Raw data).
- Links to Pubmed citations:** Points to the Citations section, which lists three PubMed entries related to the experiment.
- Links session:** Points to the Links section, which provides links to the ArrayExpress Atlas, Array design, and Experimental protocols.
- Files session:** Points to the Files section, which lists various data files such as processed data (.zip), raw data (.zip), and MAGE-TAB files (IDF, SDRF, SDF).
- List of experimental factors included in the experiment:** Points to the Experimental factors table, which lists factors like Compound, Genotype, and minutes after start of treatment, along with their values.

Fig. 12: Expanded view of a single experiment with links to several experiment annotation files.

This detailed experimental view is subdivided into the following sessions:

- Description - a description of the **experiment** supplied by the submitter;
- MIAME score - this score indicates how close to full MIAME-compliance an **experiment** is, with a score of 5 being the highest. One point is given for each of the following: array design(s), protocols, **experimental factors**, raw and processed data files;
- Contact - the name and email address of the submitter;
- Citations - publication details related to the selected **experiment**, including links to the PubMed entry, when available;
- Links to:
 - (i) **ArrayExpress Atlas of Gene Expression** showing the most differentially expressed genes in the **experiment** (if the experiment is present in the Atlas);
 - (ii) information about the array design(s) used;
 - (iii) the protocols used;
 - (iv) the ArrayExpress Advanced interface for the **experiment**. The Advanced Interface is described in more details in session 4 of this tutorial;
- Files, providing links to:
 - (i) data archives, meaning the raw and processed data files as .zip archives;
 - (ii) the **MAGE-TAB Investigation Description File (IDF)**, which provides top level information about the **experiment**;
 - (iii) the **MAGE-TAB Sample and Data Relationship File (SDRF)**, which provides information about **samples, extracts, labeled extracts, hybridizations** and the data files associated with them;

- (iv) Experiment Design Images - links to a diagram of the **sample** relationships in .png and .svg format. These graphs represent the processing steps for each **sample**, each edge represents a protocol that has been applied;
 - (v) the **MAGE-TAB** Array Design Files (ADF), describing the array designs used in the experiment;
 - (vi) the "Browse all available files" link takes you to a web page listing all the files, relating to the **experiment**, that are available to download (Fig. 13);
- **Experiment** types - terms describing experimental design types. These can include biological, methodological and technology types e.g. disease state, strain or line, compound treatment, dye swap, co-expression, etc.;
 - **Experiment factors** - a list of the **experimental factor** names and values used in the experiment;
 - **Sample** attributes - a list of the attributes used to describe the **samples**.

Experiment E-MEXP-29		
Transcription profiling of stress treated fission yeast (stressors: compound: heavy metal, oxidation, alkylation and osmosis)		
E-MEXP-29.README.txt	4 KB	23 January 2009, 16:20
E-MEXP-29.processed.zip	1.6 MB	23 January 2009, 16:20
E-MEXP-29.raw.zip	44.4 MB	23 January 2009, 16:20
E-MEXP-29.idf.txt	28 KB	23 January 2009, 16:20
E-MEXP-29.idf.xls	35 KB	23 January 2009, 16:20
E-MEXP-29.sdf.txt	127 KB	23 January 2009, 16:20
E-MEXP-29.sdf.xls	213 KB	23 January 2009, 16:20
E-MEXP-29.2columns.txt	44 KB	23 January 2009, 16:20
E-MEXP-29.2columns.xls	55 KB	23 January 2009, 16:20
E-MEXP-29.biosamples.map	70 KB	23 January 2009, 15:48
E-MEXP-29.biosamples.png	369 KB	23 January 2009, 15:48
E-MEXP-29.biosamples.svg	324 KB	23 January 2009, 15:48
E-MEXP-29.mageml.tgz	44.4 MB	15 April 2005, 11:42
Array Design A-SNGR-7		
Sanger Institute S. pombe array 2.1.1 template 2.2		
A-SNGR-7.README.txt	1 KB	6 March 2008, 18:44
A-SNGR-7.adf.txt	821 KB	7 December 2007, 13:18
A-SNGR-7.features.txt	242 KB	10 February 2006, 12:40
A-SNGR-7.mageml.tgz	463 KB	15 April 2005, 11:06
A-SNGR-7.reporters.txt	294 KB	10 February 2006, 12:40
Array Design A-SNGR-8		
Sanger Institute S. pombe array 2.2.1 template 3.2		
A-SNGR-8.README.txt	1 KB	6 March 2008, 18:44
A-SNGR-8.adf.txt	844 KB	7 December 2007, 13:18
A-SNGR-8.features.txt	242 KB	10 February 2006, 12:40
A-SNGR-8.mageml.tgz	516 KB	15 April 2005, 11:06
A-SNGR-8.reporters.txt	294 KB	10 February 2006, 12:40
Array Design A-SNGR-9		
Sanger Institute S. pombe array 3.1.1 template 4		
A-SNGR-9.README.txt	1 KB	6 March 2008, 18:44
A-SNGR-9.adf.txt	844 KB	7 December 2007, 13:18
A-SNGR-9.features.txt	242 KB	10 February 2006, 12:40
A-SNGR-9.mageml.tgz	505 KB	15 April 2005, 11:06
A-SNGR-9.reporters.txt	295 KB	10 February 2006, 12:40

Fig. 13: "Browse all available files" window. All the files related to the selected experiment are available for direct download. A README.txt file is available for each subgroup, describing the content of individual files.

4 How to query the ArrayExpress Advance Interface

Click on the 'ArrayExpress Advanced Interface' link (Fig. 12). This will bring you to the Advance Interface for experiment E-MEXP-29. This page displays the accession number for the **experiment**, the submitters name and laboratory, an automatically generated description of the **experiment**, and a free text description provided by the submitter (Fig. 14).

Various clickable options are provided to retrieve additional information (please note that most of these links are equivalent to the ones described above for the **experiment** view and will not be discussed any further):

- Retrieve data (see below);
- Experimental protocols;
- Providers, containing information about the submitters and institutions they are associated with;
- Array design(s) used;
- **Experiments** directory on the FTP server;
- **MAGE-ML**, **sample** annotation, **experiment** design, detailed **sample** annotation, data archives downloads;
- Bibliographic references;
- **Samples**, containing information about each of the original biological materials used in the study.

Fig. 14: Advance Interface for experiment E-MEXP-29

Click on the 'Retrieve data' link. The new page header provides information on the data available for the selected **experiment**.

Two data formats are available in this case:

- Processed Data Group: the normalized and/or analyzed data
- Measured Data Group: the raw data

Take a look at the 'Processed Data Group 1' (Fig. 15).

E-MEXP-29
experimental data

1 Processed Data Groups
Processed data group 1

4 Measured Data Groups
Raw data group 1
Raw data group 2
Raw data group 3
Raw data group 4

Processed data from experiment **E-MEXP-29**

Processed data group 1 Export data >>

Select any experimental condition (from 59 available), any quantitation type (from 1 available), any design element properties (from 7 available), and then press the **Export data** button to get the expression matrix.
If no design element property is selected, reporter or composite sequence names and numeric database identifiers will be exported (works faster).

59 Experimental conditions

select all inv	EXPERIMENTAL CONDITIONS	EXPERIMENTAL FACTORS			
		compound	strain or line	temperature	time
<input checked="" type="checkbox"/>	MBA.MEXP.763 >	none	wild type	39 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP.764 >	none	wild type	39 degree_C	60 m
<input checked="" type="checkbox"/>	MBA.MEXP.767 >	none	wild type	39 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP.768 >	none	wild type	39 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.771 >	none	sty1 k.o.	39 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.772 >	none	sty1 k.o.	39 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.773 >	none	sty1 k.o.	39 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.774 >	none	sty1 k.o.	39 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.775 >	none	atf1 k.o.	39 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.776 >	none	atf1 k.o.	39 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.777 >	methylmethan sulfonate	wild type	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.778 >	methylmethan sulfonate	wild type	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.781 >	methylmethan sulfonate	wild type	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.782 >	methylmethan sulfonate	wild type	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.783 >	methylmethan sulfonate	atf1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.784 >	methylmethan sulfonate	atf1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.785 >	methylmethan sulfonate	sty1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.786 >	methylmethan sulfonate	sty1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.787 >	methylmethan sulfonate	sty1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.788 >	methylmethan sulfonate	sty1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.791 >	sorbitol	wild type	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.792 >	sorbitol	wild type	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.795 >	sorbitol	wild type	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.796 >	sorbitol	wild type	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.797 >	sorbitol	atf1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.798 >	sorbitol	atf1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.799 >	sorbitol	sty1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.800 >	sorbitol	sty1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.801 >	sorbitol	sty1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.802 >	sorbitol	sty1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.803 >	cadmium sulfate	wild type	30 degree_C	0 m
<input type="checkbox"/>	MBA.MEXP.804 >	cadmium sulfate	wild type	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.805 >	cadmium sulfate	wild type	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.1325 >	cadmium sulfate	wild type	30 degree_C	0 m
<input type="checkbox"/>	MBA.MEXP.1340 >	cadmium sulfate	wild type	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.1341 >	cadmium sulfate	wild type	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.1342 >	cadmium sulfate	atf1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.1343 >	cadmium sulfate	atf1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.1344 >	cadmium sulfate	sty1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.1345 >	cadmium sulfate	sty1 k.o.	30 degree_C	60 m
<input checked="" type="checkbox"/>	MBA.MEXP.1346 >	hydrogen peroxide	wild type	30 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP.1347 >	hydrogen peroxide	wild type	30 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP.1348 >	hydrogen peroxide	wild type	30 degree_C	60 m
<input checked="" type="checkbox"/>	MBA.MEXP.1349 >	hydrogen peroxide	wild type	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.1471 >	hydrogen peroxide	atf1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.1472 >	hydrogen peroxide	atf1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.1473 >	hydrogen peroxide	sty1 k.o.	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.1474 >	hydrogen peroxide	sty1 k.o.	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.1475 >	none	none	30 degree_C	15 m
<input type="checkbox"/>	MBA.MEXP.1476 >	none	none	30 degree_C	60 m
<input type="checkbox"/>	MBA.MEXP.1477 >	none	none	30 degree_C	15 m
<input type="checkbox"/>	DBA.MEXP.219 Average.WTOXIDHYBS2 >				
<input type="checkbox"/>	DBA.MEXP.219 Average.WTOXIDHYBS1 >				
<input type="checkbox"/>	DBA.MEXP.219 Average.WTHEATHYBS1 >				
<input type="checkbox"/>	DBA.MEXP.219 Average.WTHEATHYBS2 >				
<input type="checkbox"/>	DBA.MEXP.219 Average.WTOSMOHYBS2 >				
<input type="checkbox"/>	DBA.MEXP.219 Average.WTOSMOHYBS1 >				
<input type="checkbox"/>	DBA.MEXP.219 Average.WTMMSHYBS1 >				
<input type="checkbox"/>	DBA.MEXP.219 Average.WTMMSHYBS2 >				

Fig. 15: Data retrieval page, Processed data group detail—Experimental conditions. This section of the page allows the user to select the experimental conditions to be included in the data matrix for further analysis

A list of all hybridizations (or experimental conditions) is displayed together with the corresponding **experimental factors**, in this case compound, strain, temperature and time. Each hybridization corresponds to a data file which will be used to generate a **data matrix** for analysis. The user can select all experimental conditions or only a subset. For this example we will select only 8 conditions: MEXP-763, 764, 767, 768 (wild type untreated) and MEXP-1346, 1347, 1348, 1349 (wild type treated with hydrogen peroxide), as shown using red boxes in Fig. 15. Now it is

possible to select the type of data and annotation to be exported. Scroll down to the ‘Quantitation type’ and the ‘Array Annotation’ tables (Fig. 16).

1 QuantitationType
 Protocols and softwares used:
 P-MEXP-1078 : SW:MEXP:3746
 P-MEXP-2650 : [no given software]

QUANTIFICATION TYPES

Normalized >

Array Annotation
 Array designs used:
 A-SNGR-7 A-SNGR-8 A-SNGR-9

DESIGN ELEMENT PROPERTIES

Database DB:genedb
 Feature coordinates: metaColumn metaRow column row
 Reporter control type
 Reporter group
 Reporter identifier
 Reporter name
 Reporter sequence type

[Export data >>](#)

Fig. 16: Data retrieval page, Processed data group detail— Quantitation Types and Design Element Properties. This section of the page allows the user to select the format of normalized data and the type of annotation to be included in the data matrix for further analysis

The ‘Quantitation type’ session lists all data formats available. For this **experiment**, only one quantitation type is given but for other array platforms (e.g. **Affymetrix**) more types are available. Select ‘Quantitation type: normalized’. The ‘Array annotation’ session lists the annotation information available for the array platform used. For example the user can select ‘Database DB:genedb’. At this point, the normalised data can be exported by clicking the ‘export data’ button. Before exporting the processed data, scroll down and take a look at the ‘Raw Data Group 1’. The raw data is organised in the same way as the processed data. First there is a table for all experimental conditions (Fig. 17) and then a table for Quantitation types and array Annotation (Fig. 18)

Raw data group 4

Select any experimental condition (from 61 available), any quantitation type (from 36 available), any design element properties (from 7 available), and then press the **Export data** > button to get the expression matrix.
 If no design element property is selected, reporter or composite sequence names and numeric database identifiers will be exported (works faster).

[Export data >>](#)

61 Experimental conditions

EXPERIMENTAL CONDITIONS		EXPERIMENTAL FACTORS			
select all inv		compound	strain or line	temperature	time
<input checked="" type="checkbox"/>	MBA.MEXP:761 >	none	wild type	39 degree_C	0 m
<input checked="" type="checkbox"/>	MBA.MEXP:762 >	none	wild type	39 degree_C	0 m
<input checked="" type="checkbox"/>	MBA.MEXP:763 >	none	wild type	39 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP:764 >	none	wild type	39 degree_C	60 m
<input checked="" type="checkbox"/>	MBA.MEXP:765 >	none	wild type	39 degree_C	0 m
<input checked="" type="checkbox"/>	MBA.MEXP:766 >	none	wild type	39 degree_C	0 m
<input checked="" type="checkbox"/>	MBA.MEXP:767 >	none	wild type	39 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP:768 >	none	wild type	39 degree_C	60 m
<input checked="" type="checkbox"/>	MBA.MEXP:769 >	methyImethan sulfonate	wild type	30 degree_C	0 m
<input checked="" type="checkbox"/>	MBA.MEXP:770 >	methyImethan sulfonate	wild type	30 degree_C	0 m
<input checked="" type="checkbox"/>	MBA.MEXP:771 >	none	sty1 k.o.	39 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP:772 >	none	sty1 k.o.	39 degree_C	60 m
<input checked="" type="checkbox"/>	MBA.MEXP:773 >	none	sty1 k.o.	39 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP:774 >	none	sty1 k.o.	39 degree_C	60 m
<input checked="" type="checkbox"/>	MBA.MEXP:775 >	none	atf1 k.o.	39 degree_C	15 m
<input checked="" type="checkbox"/>	MBA.MEXP:776 >	none	atf1 k.o.	39 degree_C	60 m

Fig. 17: Data retrieval page, Raw data group detail—Experimental conditions

Take a look at the ‘Quantitation types’ (Fig. 18). This **experiment** used two-colour microarrays so the data extracted from each individual feature is provided for both Cy3 (F532) and Cy5 (F635), including foreground and background intensities (mean, median and standard deviation) as well as ratio values and background corrected intensities. Any combination of these parameters can be included in the final **data matrix**, whenever raw data is needed.

36 QuantitationTypes

Protocols and softwares used:
No indication available

select	QUANTITATION TYPES
all inv	
<input type="checkbox"/>	% > B532+1SD2 >
<input type="checkbox"/>	% > B532+2SD2 >
<input type="checkbox"/>	% > B635+1SD2 >
<input type="checkbox"/>	% > B635+2SD2 >
<input type="checkbox"/>	B Pixels >
<input type="checkbox"/>	B532 Mean >
<input type="checkbox"/>	B532 Median >
<input type="checkbox"/>	B532 SD2 >
<input type="checkbox"/>	B635 Mean >
<input type="checkbox"/>	B635 Median >
<input type="checkbox"/>	B635 SD2 >
<input type="checkbox"/>	Dia. >
<input type="checkbox"/>	F Pixels >
<input type="checkbox"/>	F532 % Sat. >
<input type="checkbox"/>	F532 Mean >
<input type="checkbox"/>	F532 Mean - B532 >
<input type="checkbox"/>	F532 Median >
<input type="checkbox"/>	F532 Median - B532 >
<input type="checkbox"/>	F532 SD >
<input type="checkbox"/>	F635 % Sat. >
<input type="checkbox"/>	F635 Mean >
<input type="checkbox"/>	F635 Mean - B635 >
<input type="checkbox"/>	F635 Median >
<input type="checkbox"/>	F635 Median - B635 >
<input type="checkbox"/>	F635 SD >
<input type="checkbox"/>	Flags >
<input type="checkbox"/>	Log Ratio >
<input type="checkbox"/>	Mean of Ratios >

Array Annotation

Array designs used:
A-SNGR-8

select	DESIGN ELEMENT PROPERTIES
all inv	
<input type="checkbox"/>	Database DB:genedb
<input type="checkbox"/>	Feature coordinates: metaColumn metaRow column row
<input type="checkbox"/>	Reporter control type
<input type="checkbox"/>	Reporter group
<input type="checkbox"/>	Reporter identifier
<input type="checkbox"/>	Reporter name
<input type="checkbox"/>	Reporter sequence type

Fig. 18: Data retrieval page, Raw data group detail— Quantitation Types and Design Element Properties

According to the array platform used, different data format will be available for retrieval. Go back to 'Processed Data Group 1' and click on Export data. A **data matrix** will be computed using all selected experimental conditions, the normalised signal from each condition and the selected annotation for each identifier present on the array. On the next page, click on 'See data matrix' to view the generated file and on 'download data matrix' to save it onto your computer as .csv file (Fig. 19).

The screenshot shows a web interface for data matrix preparation. At the top, a progress bar indicates 'Preparation of data matrix 100% done'. Below this, a small table displays a data matrix with 5 rows and 5 columns of numerical values. Two buttons are visible: 'See data matrix >>' and 'Download data matrix >>', with the latter circled in red. A dialog box titled 'Opening E-MEXP-29_1504222838_1229768971_2...' is open, showing the file path and options to 'Open with Notepad (default)' or 'Save to Disk'.

Fig. 19: Data matrix download window

5 How to query the ArrayExpress Archive by experiment accession number

Any **experiment** accession number can be used to query the Archive. This is particularly useful when the user needs to retrieve a specific dataset, for example linked to a published paper of interest. In the following exercise we will use the E-MEXP-886 dataset.

1. Open the AE homepage, at <http://www.ebi.ac.uk/arrayexpress>, in a Web browser
2. In the 'Experiments' box, on the left-hand side of the page, type the **experiment** accession number E-MEXP-886 and click query (similar to what was shown in Fig. 9). This will bring up the now familiar window showing the **experiment** retrieved.
3. Expand the **experiment** view and explore the **experiment** properties. In this study, **Affymetrix** MOE430A arrays were used for transcription profiling of ataxin-null versus wild type mice to investigate spinocerebellar ataxia type 1 [4]. Explore the 'Detailed data retrieval page' for more information on the data formats available for **Affymetrix** arrays. The .CHP file contains the processed/normalized expression levels of each gene on the array and the .CEL file contains the raw data for every feature on the chip.
4. Click on the **Affymetrix** raw data icon on the title line to download onto your PC a zip archive containing all .CEL files related to this **experiment** (Fig. 20). The raw data archive just retrieved can be directly loaded and analyzed using data analysis tools such as **Expression Profiler**, an online data analysis tool provided by the EBI.

The screenshot displays the ArrayExpress interface for experiment E-MEXP-886. The 'Files' section is expanded, showing a list of files including 'E-MEXP-886.processed.zip', 'E-MEXP-886.raw.zip', 'E-MEXP-886.chp', 'E-MEXP-886.edf.txt', 'E-MEXP-886.edf.txt', 'E-MEXP-886.bioamples.png', 'E-MEXP-886.bioamples.png', and 'E-MEXP-886.wdft.txt'. A red circle highlights the 'Raw data' icon in the top right corner of the file list. A dialog box titled 'Opening E-MEXP-886.raw.zip' is open, showing the file name and options to 'Open with WinRAR.ZIP (default)' or 'Save File'.

Fig. 20: Expanded view of a single experiment – entire raw dataset download

Glossary

4DXpress

This database provides a platform to query and compare gene expression data during the development of the major model animals (zebrafish, drosophila, medaka, mouse). The high resolution expression data was acquired through whole mount in situ hybridisation-, antibody- or transgenic experiments (<http://4dx.embl.de/4DXpress/welcome.do>)

Affymetrix

Leading manufacturer of oligonucleotide arrays (<http://www.Affymetrix.com/>)

ArrayExpress Atlas of Gene Expression

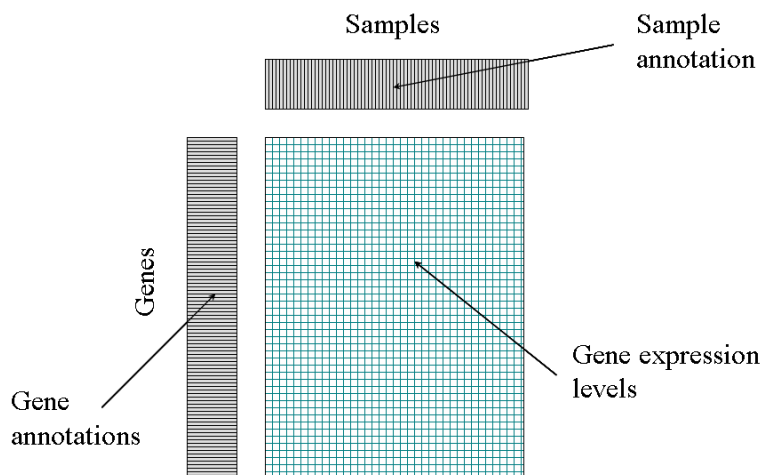
A semantically enriched database of meta-analysis based summary statistics over a curated subset of ArrayExpress Archive; it allows exploring in which conditions a gene is expressed or which genes are differentially expressed in a particular condition, tissue, cell type, etc (<http://www.ebi.ac.uk/microarray-as/atlas/>)

Bioconductor

Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data (<http://www.bioconductor.org/>)

Data matrix

In a gene expression **data matrix**, each row represents a gene and each column represents an experimental sample or array. An entry in the **data matrix** usually represents the expression level or expression ratio of a gene in a given sample or array. In addition to numerical values, the matrix can also contain additional columns for **gene annotation** or additional rows for **sample annotation**.



Ephb3

EPH-related tyrosine kinase receptor B3; receptor for members of the ephrin-b family, binds to ephrin-b1 and -b2.

Experiment

The complete set of hybridizations performed in a study.

Experiment accession number

A unique identifier assigned to each **experiment** by the AE curation staff.

Experimental factor (or Factor Value)

A property that varies between samples and it is important in the interpretation of your data (e.g. time, compound, genotype, etc.). You need a separate FactorValue[*factor*] column for each factor in the **experiment**, e.g. FactorValue[Compound]. In each FactorValue column enter the value relevant to that particular **sample**, e.g. doxycycline.

Expression Profiler

Online platform for analysis of microarray gene expression data, provided by the EBI (<http://www.ebi.ac.uk/expressionprofiler/>)

Extract

The RNA, DNA or protein extracted from a Sample. Enter a unique name for each Extract used.

Gene expression profile

A **gene expression profile** describes the (relative) expression levels of a gene across a set of experimental conditions

Gene ontology (GO)

GO is a controlled vocabulary used to describe the biology of a gene product in any organism. There are 3 independent sets of vocabularies, or ontologies, that describe: the molecular function of a gene product, the biological process in which the gene product participates and the cellular component where the gene product can be found (<http://www.geneontology.org>)

Hybridization

A single array or chip which has one or two LabeledExtracts hybridized to it.

LabeledExtract

RNA, DNA or protein labeled with a particular dye such as biotin or Cy3. You can create multiple LabeledExtracts from the same Extract but remember to give them different names if they were labeled with different dyes, e.g. 'extract A cy3', 'extract A cy5'.

MAGE-ML

MAGE-ML (Microarray Gene Expression - Markup Language) is the file format used to load experiment and array design information into ArrayExpress. After loading into ArrayExpress the original MAGE-ML file is made available for download so that it can be used for loading into other databases.

MAGE-TAB

A simple tab-delimited, spreadsheet-based format, that is used for annotating and communicating microarray data in a MIAME compliant fashion.

MIAME

Minimal information about a microarray **experiment** as recommended by the Microarray and Gene Expression Data (MGED) Society (<http://www.mged.org/>)

MINSEQE

Minimum Information about a high-throughput SeQuencing Experiment (<http://www.mged.org/minseqe/>)

Nfkbia

Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha; tumor suppressor gene.

Normalisation

Normalisation is a fundamental pre-processing step in microarray data analysis. It aims to compensate for systematic technical differences between arrays, to see more clearly the systematic biological differences between samples.

Normalisation methods

Although the aim of **normalisation** stays the same, the algorithms used for normalizing 2-color arrays differ from those used for normalizing 1-color arrays (e.g. **Affymetrix**). The MAS5 **normalisation** method is specific to **Affymetrix** gene expression data. For more information on MAS5 see http://www.Affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

OMIM

Online Mendelian Inheritance in Man; this database is a catalog of human genes and genetic disorders (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim>)

p-value

The probability of an event or outcome in a statistical experiment

Probe

A **probe** is the ssDNA molecule printed on the array.

QuickGO

QuickGO is a fast web-based browser for Gene Ontology terms and annotations, provided by the GOA group based at the EBI (<http://www.ebi.ac.uk/ego/>).

Sample

A biological material used in the study, e.g. a mouse, a tumor sample, a bacterial culture, a group of seedlings. You'll need at least one **sample** for each condition studied. If your **experiment** includes biological replicates create a **sample** for each biological replicate. If your **experiment** uses a common reference create this as a **sample** too.

Similarity search

Aims to quantify to what degree two expression profiles are similar. Such measure of similarity is called distance; the more distant two expression profiles are, in the multidimensional space, the more dissimilar they are. Distance can be measured in many different ways. In this case each **experiment** is treated separately and independently. **Experimental factors** are ignored for

similarity search purposes. For each gene, all pairwise Euclidean distances are computed (from each **probe** on the array for that gene) and the top most similar ones are returned.

Transcriptomics

Transcriptomics is the global analysis of gene expression using high-throughput technologies such as microarrays and high-throughput sequencing (HTS).

Further reading

1. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data*. Nat Genet, 2001. **29**(4): p. 365-71.
2. Brazma, A., et al., *ArrayExpress--a public Archive for microarray gene expression data at the EBI*. Nucleic Acids Res, 2003. **31**(1): p. 68-71.
3. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Stat Appl Genet Mol Biol, 2004. **3**(12): p. Article3.
4. Goold, R., et al., *Down-regulation of the Dopamine Receptor D2 in mice lacking Ataxin 1*. Hum Mol Genet, 2007. **28**: p. 28.