

ArrayExpress and Expression Atlas practical: querying and exploring gene expression data at EMBL-EBI

Last updated: 28 August 2013

Amy Tang – amytang@ebi.ac.uk

This practical will introduce you to the data content and query functionality of ArrayExpress and Gene Expression Atlas databases. We suggest using Firefox for this tutorial.

Additional information on these two resources including dedicated courses and more exercises can be found on the EBI eLearning portal, Train Online.

ArrayExpress: <http://www.ebi.ac.uk/training/online/course/arrayexpress-exploring-functional-genomics-data-ar>

Expression Atlas: <http://www.ebi.ac.uk/training/online/course/arrayexpress-investigating-gene-expression-pattern-0>

Please consider that the results you will obtain while doing the exercises might differ from what illustrated here due to a recent database update.

ArrayExpress Exercise 1

- Imagine you found an ArrayExpress accession in a Toxicology journal publication...

Scenario

In your literature search, you came across this paper about the effect of marine contaminated sediments on young sea breams:

Biological effects of marine contaminated sediments on Sparus aurata juveniles.
C. Ribecco, M.E. Baker, R. Šášík, Y. Zuo, G. Hardiman, O. Carnevali. *Aquatic Toxicology* (2011)

In the paper's Materials and Methods section, it quotes ArrayExpress accession E-TABM-1146.

Task

Find experiment E-TABM-1146 in ArrayExpress database (<http://www.ebi.ac.uk/arrayexpress/experiments/browse.html>) and explore the information provided. E.g. what is/are the experimental factor(s)? How would you download raw data files from the control samples only?

ArrayExpress Exercise 2

- High-throughput sequencing example

Scenario

High-throughput sequencing (HTS) is becoming a popular tool in cancer research to decipher the genetic make-up of a tumour. Mutations, epigenetic misregulation and genomic reorganisation are just some of the things that can be studied using this technology.

Imagine that you have just started a project working on human “prostate adenocarcinoma” and you want to find out all RNA-seq experiments in ArrayExpress that study this cancer.

Task

Use the ArrayExpress database (<http://www.ebi.ac.uk/arrayexpress/experiments/browse.html>) to find relevant experiments.

Other scenarios to try in ArrayExpress

- Find DNA-based experiments (e.g. ChIP-chip, genotyping, comparative genomic hybridisation) which were directly submitted to ArrayExpress, studying the effect of estrogen on humans.
-

Atlas Exercise 1

- Cigarette smoke and male (in)fertility

Scenario

Smoking is known to reduce male fertility. A recent study published in *Reproductive Toxicology* has reported dose-dependent upregulation of several genes in mouse spermatocytes which were treated with cigarette smoke condensate, including *Sod1*, *Sod2*, *Ahr*, *Arnt* and others: Esakky et al. (2012) [Cigarette smoke condensate induces aryl hydrocarbon receptor-dependent changes in gene expression in spermatocytes](#). *Reprod Toxicol.* 2012 Dec;34(4):665-76. doi: 10.1016/j.reprotox.2012.10.005. Epub 2012 Oct 13.

Imagine you’re studying male (in)fertility and would like to see if the genes reported are also expressed in human testis (as an indication of whether they may be involved in spermatogenesis).

Task

Use the Expression Atlas database (<http://www.ebi.ac.uk/gxa/>) to search for *SOD1*, *SOD2*, *AHR* or *ARNT* genes in human and see if they are expressed in human testis also. If not, why not?

According to data in the Atlas, find genes which are upregulated in human testis and are known to be associated with the function “gamete generation” (i.e. spermatogenesis).

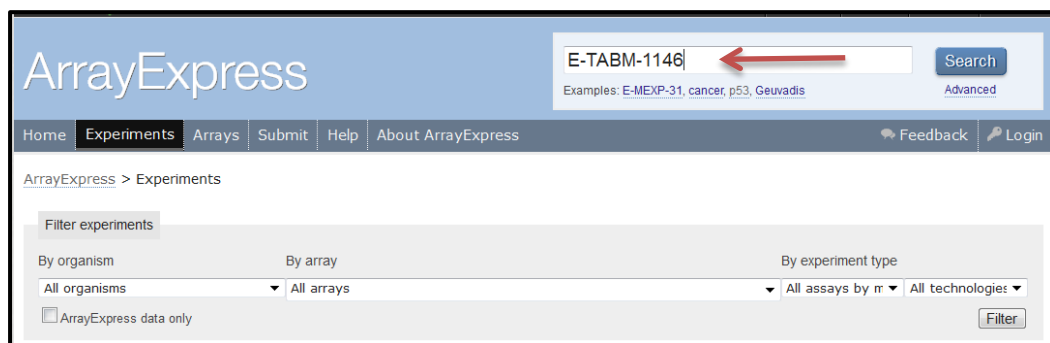
Other scenarios to try in Expression Atlas

- Find genes differentially expressed in human ovarian cancer cell lines. These genes may be molecular signatures/biomarkers for clinical diagnosis.
 - Find genes which are upregulated in human under the condition of “ultraviolet light”. Among the genes returned, search for those which have the function “DNA repair”.
-

Need some help?

ArrayExpress Exercise1

1. Open the ArrayExpress experiments page, <http://www.ebi.ac.uk/arrayexpress/experiments/browse.html>.
2. Type “E-TABM-1146” in the search box and then click “Search”:



3. The search results take you directly to a single page for E-TABM-1146. For the experimental factor (compound) studied, the “Description” field gives you some idea. To see exactly which sample is associated with which compound, go to the “Samples (3) [Click for detailed sample information and links to data](#)” link:

E-TABM-1146 - Biological effects of marine contaminated sediments on Sparus aurata juveniles	
Status	Released on 1 January 2011, last updated on 26 May 2011
Organism	Sparus aurata
Samples (3)	Click for detailed sample information and links to data
Array (1)	A-MEXP-2060 - UCSD Sparus aurata Array
Protocols (7)	Click for detailed protocol information
Description	Chemical analysis of the compounds present in sediment, although informative, often is not indicative of the downstream biological effects that these contaminants exert on resident aquatic organisms. More direct molecular methods are needed to determine if marine life is affected by exposure to sediments. In this study, we used an aquatic multispecies microarray and q-PCR to investigate the effects on gene expression in juvenile sea bream (<i>Sparus aurata</i>) of two contaminated sediments defined as sediment 1 and 2 respectively, from marine areas in Northern Italy.
Experiment types	transcription profiling by array, compound treatment
Contact	Gary Hardiman
Citation	Biological effects of marine contaminated sediments on <i>Sparus aurata</i> juveniles. C. Ribecco, M.E. Baker, R. Šášík, Y. Zhu, G. Hardiman, G. Casanova. <i>Aquatic Toxicology</i> . (2011)

E-TABM-1146 - Biological effects of marine contaminated sediments on Sparus aurata juveniles		Sample Characteristics				Factor Values		Links to Data	
Source Name	BioSourceProvider	BioSourceType	DevelopmentalStage	OrganismPart	Organism	Sex	Compound	Raw	Processed
Liver 1	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Control	↓	↓
Liver 1	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 1	↓	↓
Liver 2	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 1	↓	↓
Liver 3	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 2	↓	↓
Liver 3	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 2	↓	↓

4. On the “Samples” page, you will also find direct links to raw data file for the two control samples:

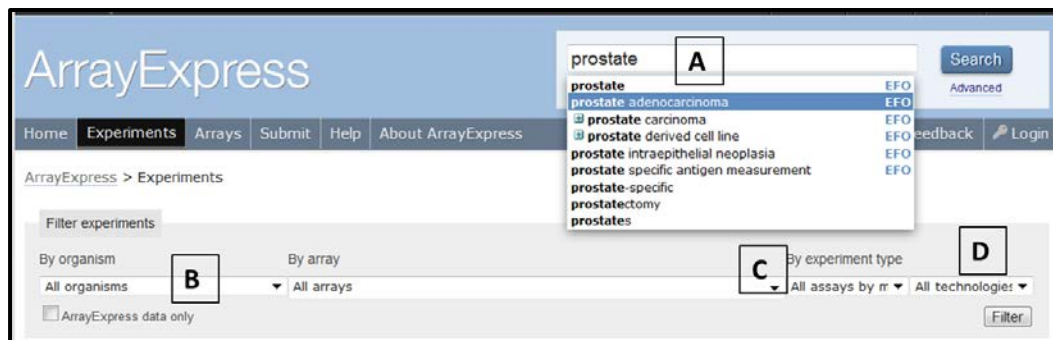
ArrayExpress > Experiments > E-TABM-1146 > Samples and Data*

E-TABM-1146 - Biological effects of marine contaminated sediments on Sparus aurata juveniles

Source Name ^	BioSourceProvider	Sample Characteristics				Factor Values		Links to Data	
		BioSourceType	DevelopmentalStage	OrganismPart	Organism	Sex	Compound	Raw	Processed
Liver 1	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Control	↓	↓
Liver 1	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Control	↓	↓
Liver 2	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 1	↓	↓
Liver 2	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 1	↓	↓
Liver 3	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 2	↓	↓
Liver 3	University of California at San Diego (BIOGEM)	wild type liver	adult	liver	Sparus aurata	male	Sediment 2	↓	↓

ArrayExpress Exercise 2

1. Open the ArrayExpress experiments page, <http://www.ebi.ac.uk/arrayexpress/experiments/browse.html> so you can see both the search box and the drop-down filters.
2. Start typing “prostate” in the experiment search box [A] and then select the EFO term “prostate adenocarcinoma”. Click “Search”.



3. Filter your search by organism “Homo sapiens” [B], “RNA assay” [C] and “Sequencing assay” [D], then click “Filter”. You do *not* need to touch the “All arrays” option as it is only used when you want to filter for experiments done on a specific microarray platform (e.g. Affymetrix mouse 3’ IVT array).
4. The results now show only prostate adenocarcinoma RNA-seq experiments in human.

The screenshot shows the ArrayExpress search results page. The search box contains the text 'prostate adenocarcinoma'. The filter options are: 'By organism' [B] set to 'Homo sapiens', 'By array' [C] set to 'All arrays', and 'By experiment type' [D] set to 'RNA assay' and 'Sequencing assay'. A 'Filter' button is located at the bottom right of the filter section. The results table shows 22 experiments.

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-GEOD-47806	LncRNA-dependent mechanisms of androgen receptor-regulated gene activation programs [GRO-seq II]	RNA-seq of coding RNA	Homo sapiens	6	15/08/2013		-	-
E-GEOD-47805	LncRNA-dependent mechanisms of androgen receptor-regulated gene activation programs [GRO-seq I]	RNA-seq of coding RNA	Homo sapiens	4	15/08/2013		-	-
E-GEOD-48230	Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer	RNA-seq of coding RNA	Homo sapiens	10	01/08/2013		-	-

5. You can now look more carefully at individual experiments to identify the one that might be more relevant for your research. Let’s explore E-GEOD-24284.

Click on experiment E-GEOD-24284 and explore the information available for this experiment.

E-GEOD-24284 - Deep sequencing analysis of transcription-induced chimeras in human prostate adenocarcinoma and reference samples

Status	Released on 22 September 2010, last updated on 6 November 2012
Organism	<i>Homo sapiens</i>
Samples (20)	Click for detailed sample information and links to data
Array (1)	A-MEXP-1597 - Agilent Human Genome CGH Microarray 244A 014693
Protocols (8)	Click for detailed protocol information
Description	This SuperSeries is composed of the following subset Series: GSE24282: CGH microarray analysis of human prostate adenocarcinoma and normal samples GSE24283: Deep transcriptional sequencing analysis of human prostate adenocarcinoma and reference samples Refer to individual Series
Experiment types	comparative genomic hybridization by array, RNA-seq of coding RNA
Contact	Thomas D. Wu <twu@gene.com>
MINSEQE	Exp. design Protocols Factors Processed Seq. reads
MIAME	Platforms Protocols Factors Processed Raw
Files	Data Archives E-GEOD-24284.processed.1.zip , E-GEOD-24284.raw.1.zip , E-GEOD-24284.raw.2.zip , E-GEOD-24284.raw.3.zip Investigation Description E-GEOD-24284.idf.txt Sample and Data Relationship E-GEOD-24284.hyb.sdrf.txt , E-GEOD-24284.seq.sdrf.txt Array Design A-MEXP-1597.adf.txt R ExpressionSet E-GEOD-24284.eSet.r Browse all available files
Links	GEO - GSE24284 Send E-GEOD-24284 data to GENOMESPACE

This is an interesting experiment because the investigator performed both microarray and sequencing analyses on the samples. In particular, take a look at the “Samples (20) Click for detailed sample information and links to data” section (circled in red in the screenshot above) to find out more information about the samples analyzed in this experiment and how they relate to the data files.

Atlas Exercise

1. Open the Atlas homepage, <http://www.ebi.ac.uk/gxa/>.
2. To search for individual genes (*SOD1* is shown here as an example), type in the “Genes” search box “SOD1” and select, from the suggestions, the matching term for the human gene. Always select one with the Ensembl gene ID shown (prefixed by “ENS”), and then click “Search Atlas”.

The screenshot shows the Atlas search interface. The search box contains 'SOD1'. Below it, a dropdown menu lists several gene suggestions. A red arrow points to the first suggestion: 'gene: SOD1 (ENSG00000142168,ALS,ALS1,IPOA) Homo sapiens (1)'. Other suggestions include SOD1P1, SOD1P2, and SOD1 from other organisms like Rattus norvegicus, Macaca mulatta, and Equus caballus. The interface also shows a 'Search Atlas' button and a 'hide suggestions' link. On the right, there is a link to 'advanced search' and a description of the Atlas database.

3. Now you have the gene view page for *SOD1*. You can scroll down the page and explore. For example, have a look at the anatomogram to see experiments in which the gene is differentially expressed.

To see the full list of organism parts, click at the link below the anatomogram:

organism part
studied in E-GEOD-15765, E-GEOD-13911, E-MEXP-1251, E-MTAB-69, E-GEOD-1563, ... (21 experiments)

Number of published studies where the gene is over/under expressed compared to the gene's overall mean expression level in the study.

Organism Part	Over (Red)	Under (Blue)
brain	3	2
craniofacial tissue	2	0
lung	1	0
heart	1	0
spinal cord	1	0
digestive system component	3	5
liver and biliary system	3	0
renal system	2	1

[show expression data for all values of this factor>>](#)

- To look for spermatogenesis genes upregulated in human testis, go back to the Atlas homepage, <http://www.ebi.ac.uk/gxa/> and start a new search.
- Start typing “gamete” in the “Genes” search box [A] and select the suggestion “gamete generation” from the drop-down list.

Atlas Release:

data	13.07
software	2.0.21.2
new experiments	146
total experiments	2841
total genes	886420

Gene Expression Atlas
The Gene Expression Atlas is a semantically enriched database of meta-analysis based summary statistics over a curated subset of ArrayExpress Archive, servicing queries for condition-specific gene expression patterns as well as broader exploratory searches for biologically interesting genes/samples.

- Restrict your search to upregulated genes and organism “Homo sapiens” [B].
- Type “testis” in the condition search box [C]. You will see a list of suggested terms in a drop-down list. Hover your mouse cursor over the suggested terms to reveal their EFO IDs. The term you need is “testis”, [EFO_0000984].

Gene Expression Atlas
The Gene Expression Atlas is a semantically enriched database of meta-analysis based summary statistics over a curated subset of ArrayExpress Archive, servicing queries for condition-specific gene expression patterns as well as broader exploratory searches for biologically interesting genes/samples.

- Click “Search Atlas” [D]. You will get a results page which looks like this:

Genes 1-32 of 32 total found (you can [refine your query](#)) • [Download all results](#)

Legend: 3 1 - number of studies the gene is **over/under** expressed in (~ in experiment pop-ups indicates non-differential expression)

Gene	Over	Under
DIAPH2	2	1
DYNLL1	2	0
ZNF148	1	1

9. The results of your search are presented as a heatmap view that shows genes with the function of “gamete generation” and were identified as upregulated in human testis. If you would like to expand your search by looking at the male reproductive system as a whole, you can expand the EFO “tree”:

