EMBL-EBI Tutorial

Gabriella Rustici (v5, 24/09/2009)

ArrayExpress Atlas for beginners

The ArrayExpress (AE) database is a public curated repository of **transcriptomics** data [1]. As of September 2009 the database holds over 9000 separate studies.

The ArrayExpress Atlas is a semantically enriched database of meta-analysis based summary statistics over a curated subset of AE data. It allows exploring in which conditions a gene is expressed or which genes are **differentially expressed** in a particular condition, tissue, cell type, etc. As of September 2009, the Atlas holds over 1100 experiments, corresponding to > 5800 conditions.

This tutorial assumes that you have completed the "ArrayExpress for Beginners" tutorial.

You will learn about:

- The ArrayExpress Atlas
- How to query the Atlas

Contents:

1 What is the AE Atlas and how to use it

The AE Atlas provides a statistically robust framework for integrating **gene expression** data across different platforms at a **meta-analytical** level. A simple interface allows the user to identify **differentially expressed** genes in conditions of interest, providing both gene and condition queries.

The Atlas uses data from the AE Warehouse which have been carefully re-annotated for comparability and have had their **gene annotation** updated using Ensembl (http://www.ensembl.org) or Uniprot (http://www.uniprot.org) . This means that information such as GO terms, **HUGO names** (http://www.genenames.org/) and other functional annotation have been added, when available. In addition, we are building an **experimental factor** ontology, or **EFO** (http://www.ebi.ac.uk/microarray-srv/efo), and using it to structure the data annotations in the Atlas. This ontology is used to enhance Atlas queries, through the EBI Ontology Look-Up service (http://www.ebi.ac.uk/ontology-lookup/).

Currently, 3700 **experiment**s from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) have been imported into AE; therefore the Atlas also provides a way to integrate gene expression data from the two databases.

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.

For constructing the Atlas, the data is taken as normalized by the submitters/authors. For every **experiment**, the strength of differential expression of genes across conditions across **experiment**s is computed and **meta-analytical** statistics across **experiment**s are calculated. The statistical models and computations are performed primarily with the aid of the **Bioconductor** package LIMMA [2].

The resulting AE Atlas presents itself as a two-dimensional matrix, akin to a **data matrix**: one dimension (rows) corresponds to genes, and the second dimension (columns) corresponds to conditions, rather than samples. The matrix entries are ***p*-value**s together with a sign, indicating the significance and direction of differential expression.

The AE Atlas user interface can be found at http://www.ebi.ac.uk/microarray-as/atlas/.

2 How to query the AE Atlas by gene

The Atlas can be queried by gene(s), direction of differential expression, organism and **experimental condition(s)** (Fig. 1). For additional information on all query options go to http://www.ebi.ac.uk/microarray/doc/atlas/faq.html#faq_list_expt_conds.

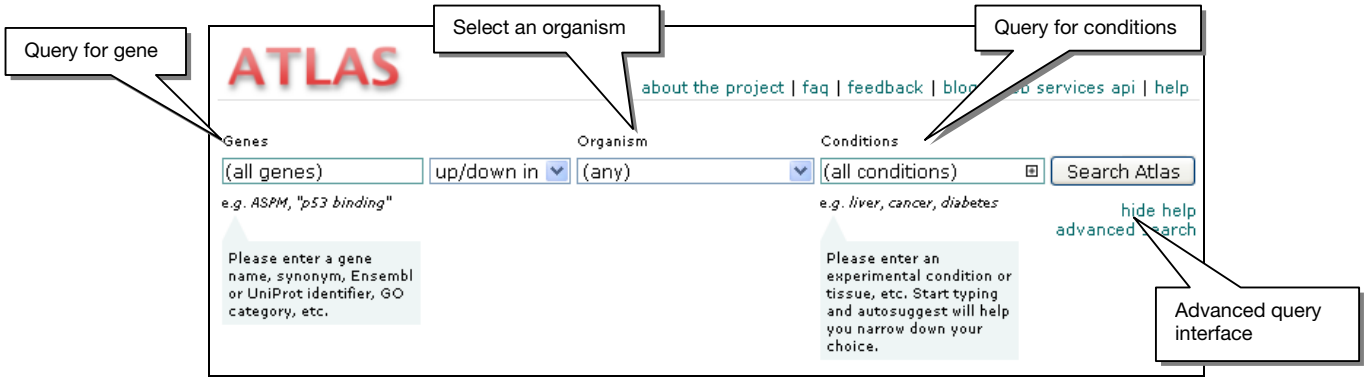


Fig. 1: The Atlas query interface (http://www.ebi.ac.uk/microarray-as/atlas/)

More specifically, querying for gene(s) will identify all genes whose annotation matches your query. Your search can be restricted only to genes belonging to a given organism and also by direction of differential expression: only up-regulated ('up in'), only down-regulated ('down in') or both ('up/down in'). The 'Conditions' parameter will identify all **experiment**s in which the conditions that match your query appear.

Any of the genes, organism or conditions boxes can be left blank but not all three. You can leave the 'Conditions' box blank. As result, the found genes will be ranked for differential expression across all conditions in the Atlas. The user might choose to specify a particular condition of interest and leave the 'Genes' box blank instead. As result, your condition query will be applied to all genes in the Atlas. We will see an example of 'condition query' in part 2 of this tutorial.

The 'Genes' and 'Conditions' boxes have auto-complete functions. Start typing and a list of suggested terms will appear. To select one of the suggested terms click on it. If you do not want to select any, click on the 'hide suggestions' link at the bottom of the list (Fig.2). The 'Conditions' box can also be filled by selecting terms of interest from the ontology tree; you can navigate the ontology tree, down to the term of interest, by clicking on the plus sign inside the box/next to each individual node of the tree.

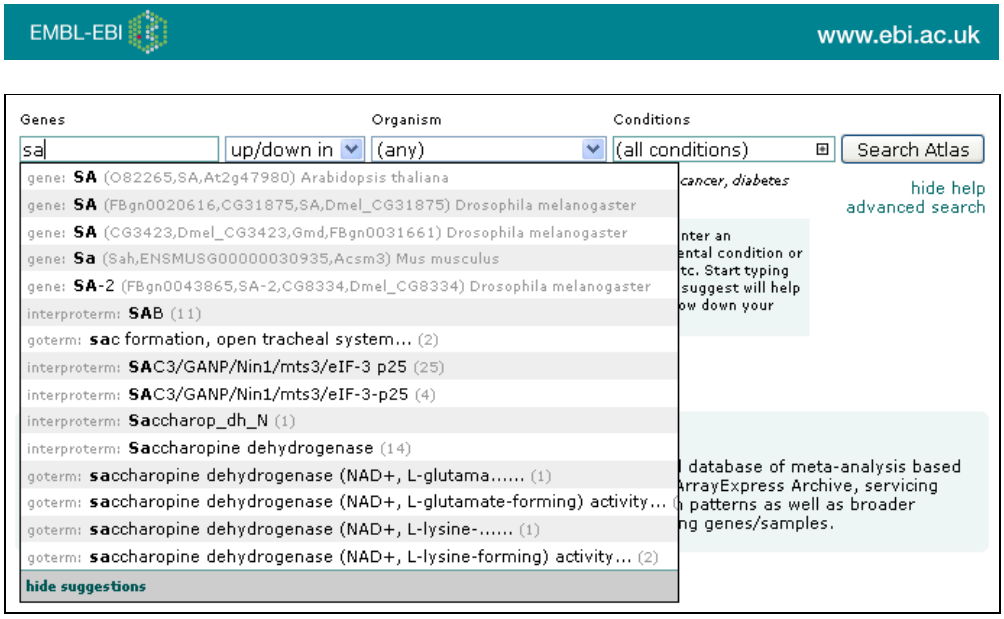


Fig. 2: An auto-complete feature example. The 'Genes' auto-complete function groups the suggested terms together first by gene and then lists general textual matches in alphabetical order (categories include, Interpro term, GO term, disease, etc.).

To familiarise you with the Atlas query interface, we will first perform a simple gene-condition query.

1. Open the Atlas homepage, at http://www.ebi.ac.uk/microarray-as/atlas/, in a Web browser
2. In the 'Genes' box, on the left-hand side of the page, type the following gene names, e.g. **saa4 aspm** (as shown in Fig. 3).
3. In the direction of differential expression drop down menu leave the default 'up/down in' selected
4. Select *Mus musculus* in the 'Organism' dropdown menu
5. In the 'Conditions' box, on the right-hand side of the page, type the following conditions, e.g kidney liver
6. Click the 'Search Atlas' button.

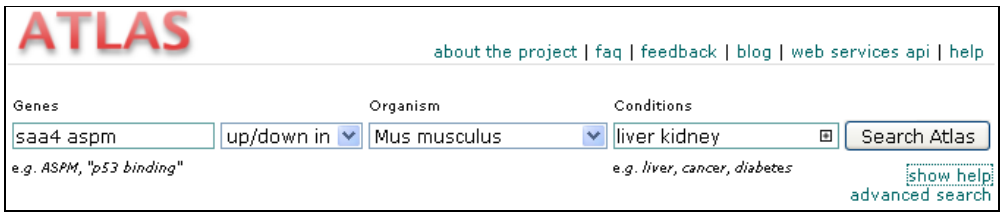


Fig. 3: The Atlas query interface (http://www.ebi.ac.uk/microarray-as/atlas/); example of a 'gene/condition' query

The results are displayed in a heatmap (Fig. 4); all the genes found are listed in the first column of the heatmap, the conditions go across the columns and the heatmap entries provide the number of times a gene was observed significantly up- or down-regulated in the corresponding condition. The heatmap cell color ranges from red, i.e., up-regulated, to blue, i.e., down-regulated. The

conditions retrieved are shown as nodes in the EFO tree (see Ontology section of the heatmap) as well as individual keywords (see Keywords section of the heatmap).

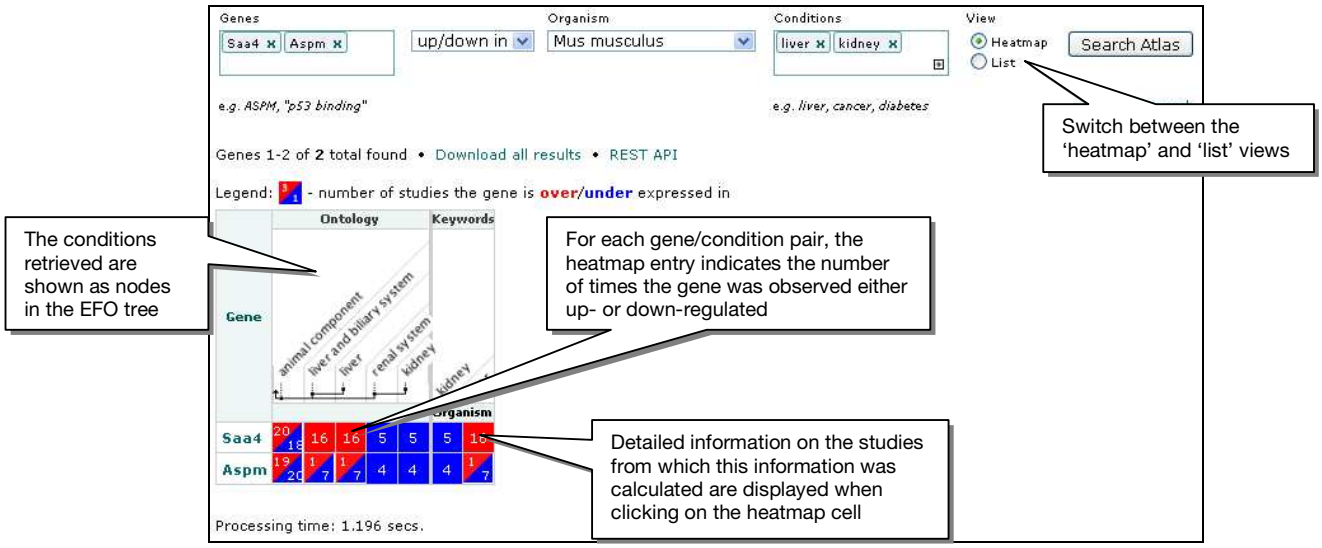


Fig. 4: Atlas heatmap output. Results can be downloaded into a tab-delimited format by clicking on "Download all results". A pop up window will open showing the progress of all download requests. Once download is complete click on "Get file" to store the file on your computer. You can also retrieve the same results using REST API.

Take a look at the EFO ontology tree; we see that the mouse gene **saa4** was observed up-regulated 16 times in liver samples, and down-regulated 5 times in kidney samples. **Aspm** instead was observed down-regulated 7 times/up-regulated 1 time in liver and 4 times in kidney.

Clicking on a cell in the heatmap will show more information about the expression of the selected gene in the associated experiments. For example, click on the cell with the number 16 in it, corresponding to **saa4** expression in liver samples.

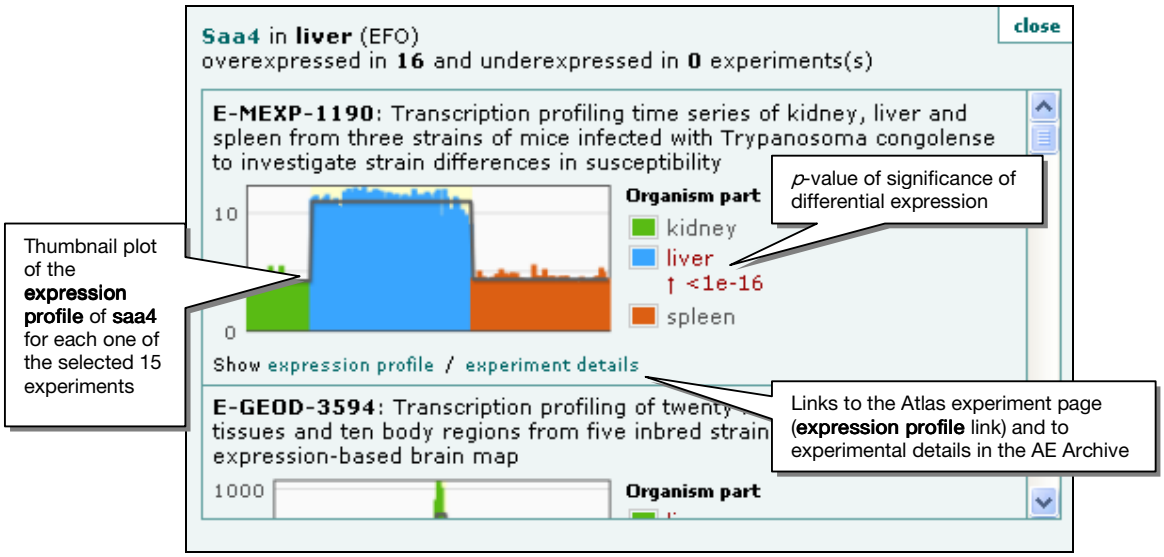


Fig. 5: Detailed view showing the over-expression of saa4 in liver samples in 16 experiments.

The window that opens up (Fig. 5) contains detailed information on the 16 experiments in which **saa4** was observed to be up-regulated in liver samples. A thumbnail plot of the **expression profile** of **saa4** in each one of the selected 16 studies is showing, with links to the **expression profile** in the Atlas experiment page and to experimental details in the AE Archive (Fig. 5). For each experiment, a brief description and the *p*-value of significance of differential expression are also displayed.

Click on the '**expression profile**' link under the thumbnail plot of the **expression profile** of saa4 for experiment E-MEXP-1190 and explore the experiment page.

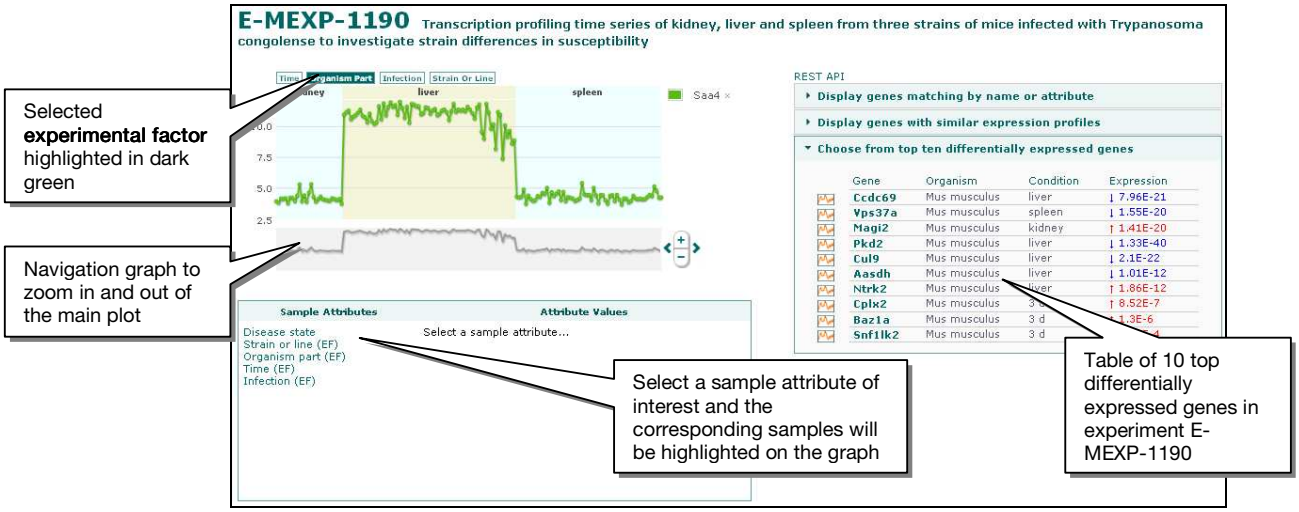


Fig. 6: Experiment page for E-MEXP-1190 showing a full size plot of the saa4 expression profile

In the graph now showing (Fig. 6), the X axis represents all samples in this study, grouped by **experimental factor** and the Y axis the expression levels for **saa4** in each **sample**. Explore the dependency of the expression levels on different **experimental factor**s. **Experimental factor**s are the main experimental variables studied in a particular study. For instance, **experiment** E-MEXP-1190 has four **experimental factor**s – 'time', 'organism part', 'infection' and 'strain or line' – which are listed above the graph. The **experimental factor** 'organism part' is now selected and you can observe that, under this condition, **saa4** has notably higher expression values in liver than in kidney or spleen (Fig. 6). You can select a different experimental factor to see if this correlation is preserved or lost.
Sample attributes can be selected from the 'Sample attributes' table. Once a particular sample attribute is selected, the related samples will be highlighted on the graph.

On the right hand side of the experiment page, several 'search' options allow you to retrieve gene(s) of interest and add its/their **expression profile**(s) to the plot.

The options available are:

- 'Display genes matching by name or attribute', which allows you to search for a gene of interest by gene identifier/attribute;
- 'Display genes with similar **expression profile**s', which allows you to search, for any of the genes currently plotted, for the 10 most similar genes (based on Pearson correlation); and

- 'Choose from top 10 differentially expressed genes', which provides a list of the 10 most differentially expressed genes, in the selected study. For each gene a p-value of significance of differential expression is provided (Fig. 7).

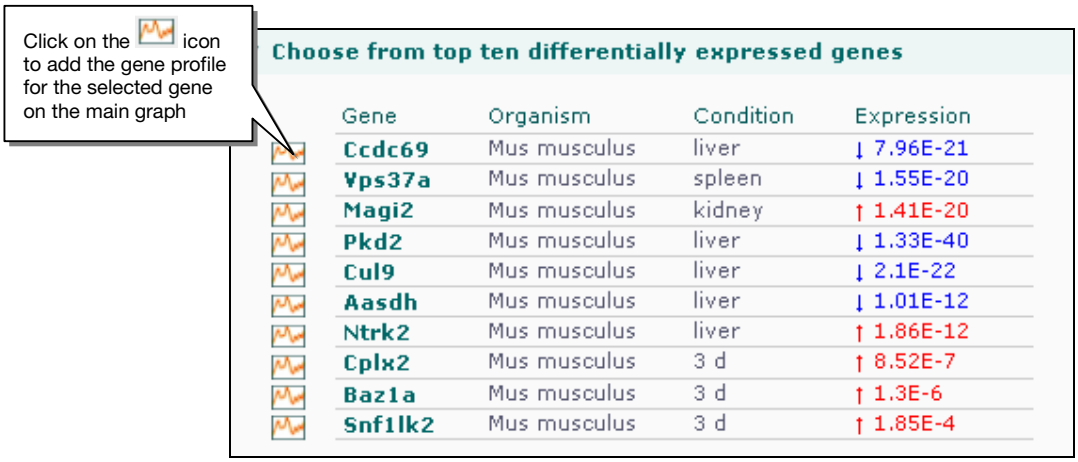


Fig. 7: Table of top 10 differentially expressed genes in E-MEXP-1190

For example, click on 'Display genes matching by name or attribute' and search for **aspm**. One gene is retrieved, matching the search criteria, and can be added to the current plot by clicking on the graph icon to the left of the gene identifier for **aspm** (Fig. 8).

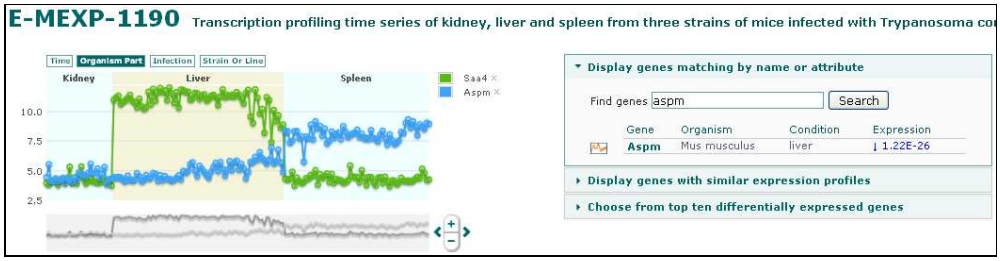


Fig. 8: Experiment page for E-MEXP-1190 showing a full size plot of the saa4 expression profile. Aspm was added to the plot using the 'Display genes matching by name' option.

Let's now take a look at the list output. Go back to the window displayed in Fig. 4 and select the 'list' view option at the top of the page, as shown in Fig. 9. Click on 'Search Atlas'.

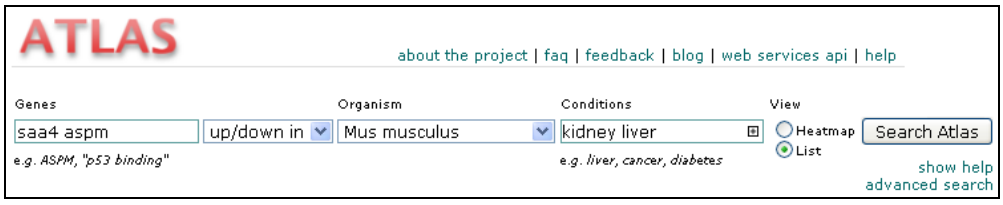


Fig. 9: The Atlas query interface (http://www.ebi.ac.uk/microarray-as/atlas/)

In the list output (Fig. 10), each row of the results corresponds to one gene/condition pair; the number of times the selected gene was observed significantly up- or down-regulated in the selected condition, together with a *p*-value of significance are provided, similarly to what described for the heatmap view. Results can be downloaded by clicking on the 'download results link'.

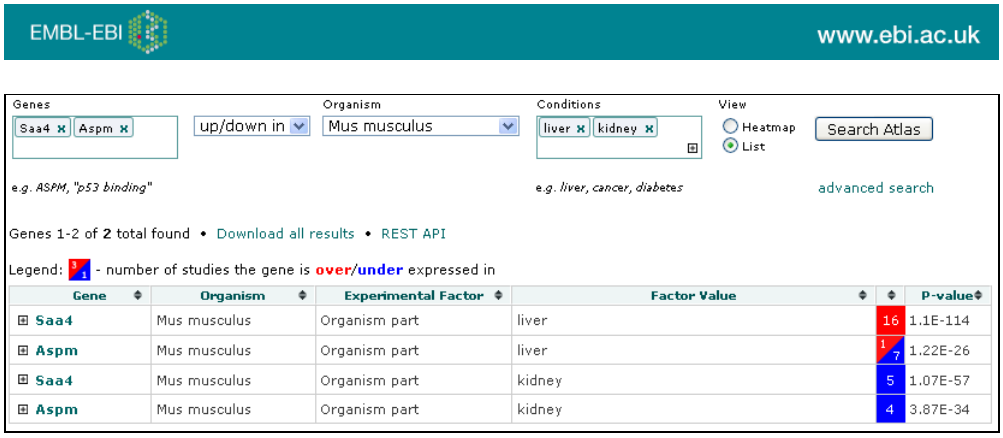


Fig. 10: Atlas list output

You can expand each row by clicking on the + sign to the left of the gene identifier. The full list of experiments in which the selected gene was observed to be differentially expressed is provided (Fig. 11).

Clicking on an experiment identifier will allow retrieving information for the selected experiment in the AE Repository. Clicking on the experiment thumbnail will bring you to the experiment page. This can be navigated as previously explained.

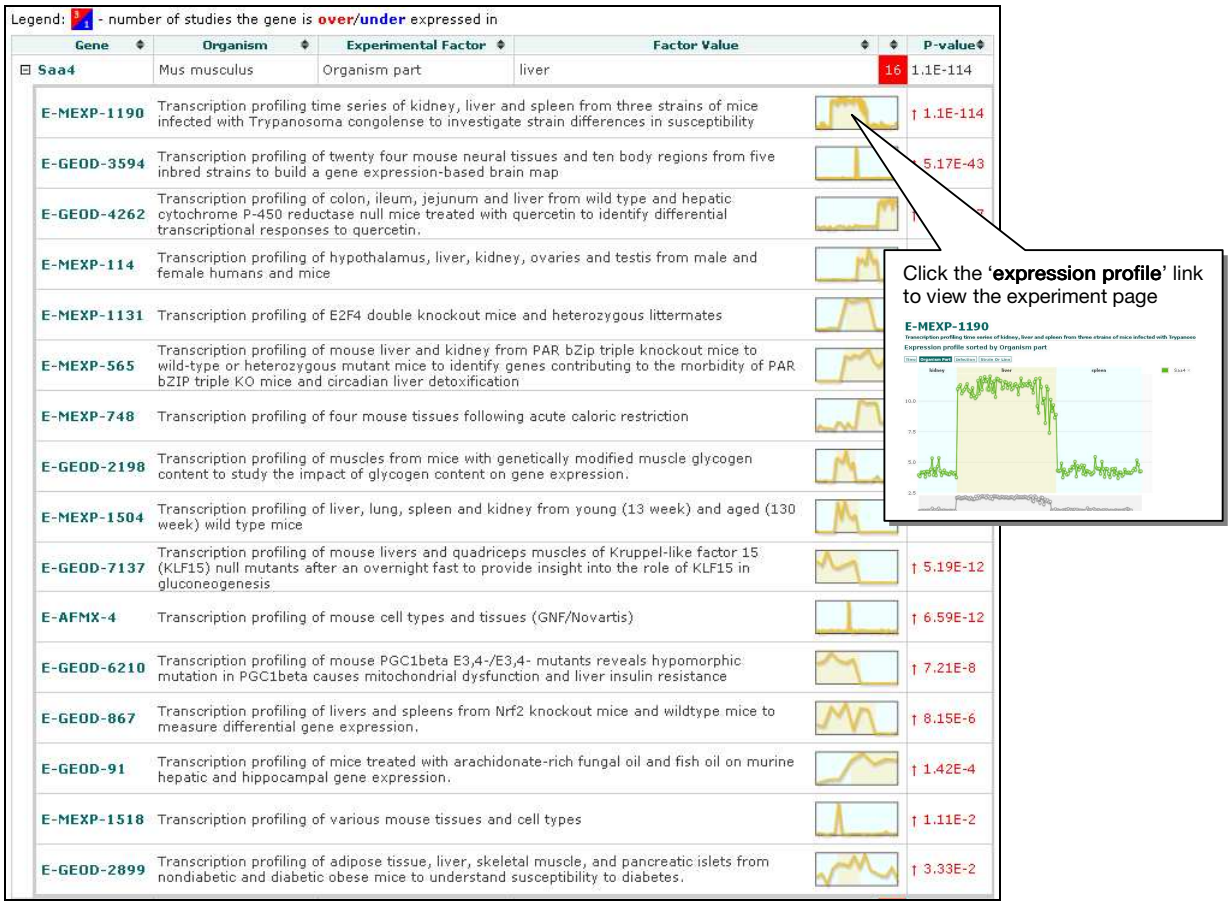


Fig. 11: Atlas list output, expanded view

3 The AE Atlas gene view page

If your search identifies only one gene, or you click on one gene identifier in the heatmap or list view, you will be redirected to the gene view page for that particular gene.

From the heatmap in Fig. 4, click on the mouse **saa4** gene name. The gene view page (Fig. 12) provides a one page summary of all the information present in the Atlas for a given gene.

The following information is displayed:

- summary of terms and external databases cross-references, including synonyms, links to InterPro and GO terms, Uniprot and Ensembl identifiers as well as **ortholog** genes, which allows comparing **ortholog**s behaviour across the Atlas. Click on the 'Compare orthologs' link to generate a heatmap of **ortholog**s to the selected gene;
- expression heat map listing all the conditions in which the gene was observed differentially expressed; and
- thumbnail images of **gene expression profile**s for the studies in which the gene was found to be differentially expressed.

Fig. 12: Gene view page for Mus musculus saa4

The mouse gene **Saa4** shows the highest level of differential expression in the experiment E-MEXP-1190 and it is up-regulated in liver and down-regulated in kidney.

Click on any of the conditions in the expression summary heatmap and you will display only the experiments where this gene is up- or down-regulated in the selected condition.

4 How to query the AE Atlas by condition

We will now perform a simple search querying for a single condition (Fig. 13).

1. Open the Atlas homepage, at http://www.ebi.ac.uk/microarray-as/atlas/, in a Web browser
2. Leave the 'Genes' box, blank
3. Leave the default 'up/down in' selected, to include both up- and down-regulated genes
4. Select species, e.g. *Homo sapiens*, in the 'Organism' dropdown menu
5. Browse to the term 'carcinoma, EFO_0000313', using the EFO ontology tree, in the 'Conditions' box and click the 'Search Atlas' button.

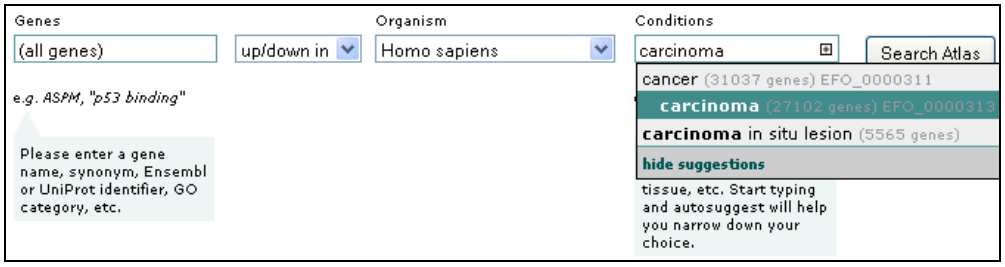


Fig. 13: The Atlas query interface (http://www.ebi.ac.uk/microarray-as/atlas/); example of a condition query

When querying for a term like 'carcinoma', your original query will be automatically expanded using the **EFO** ontology. The search will now include all child terms available for your original query making the output of your search more comprehensive (Fig. 14).

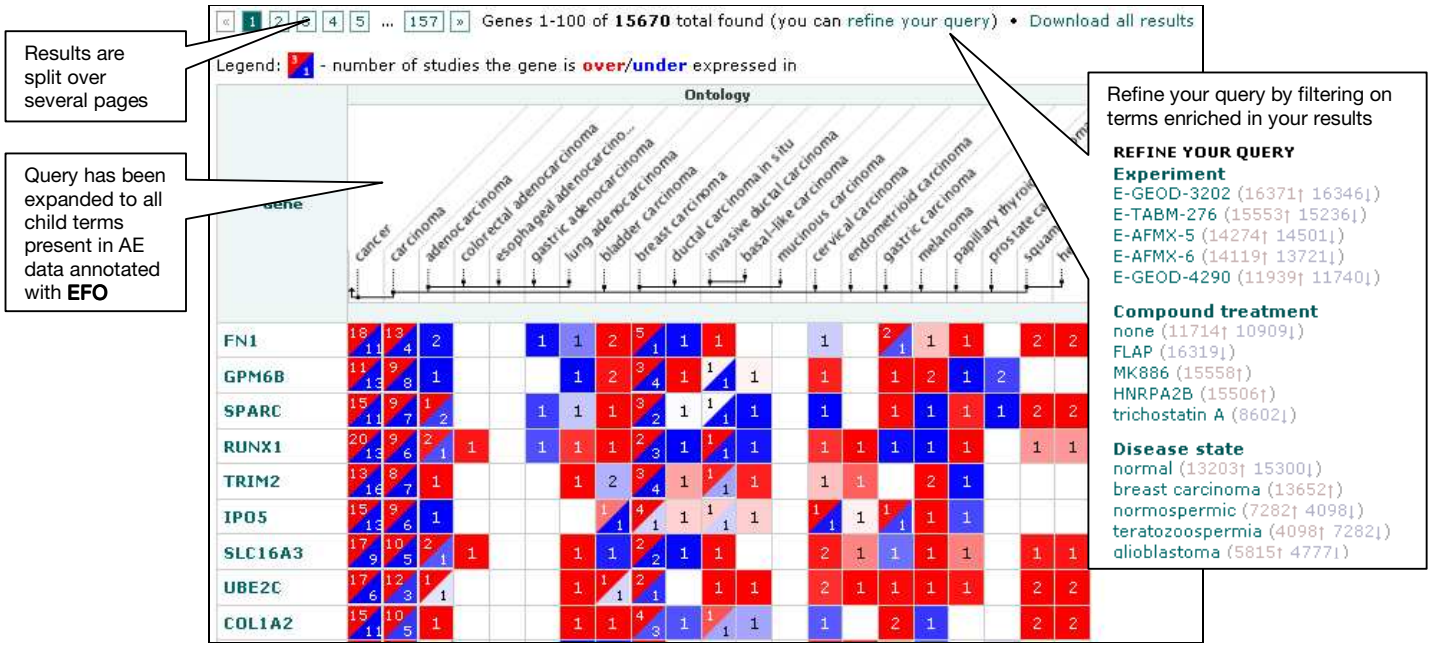


Fig. 14: Example of a condition query using the term 'carcinoma'. The query was expanded using EFO.

In this case, many genes/conditions are returned and understanding the results can be a daunting task. Filters are available on the left hand side of the heatmap to help you refine your search, by filtering on terms which are enriched in your results. Click on the 'Refine your query' link to see a list of enriched terms that can be used for filtering. Terms are grouped by category (e.g. 'Compound treatment', 'Organism part', 'Disease state', etc.); for each category, the five most frequently occurring terms are listed.

For example, we could filter the results using the terms 'Leukemia, acute myeloid' and then 'Chromosomal rearrangement'. This reduces the number of genes retrieved to just 14. The filters in use are now listed at the top of the page, below the parameters used in the starting query (Fig. 15). Additional filtering conditions can also be added to the query using the available drop down menus: 'Gene property', 'Experimental factor' and 'Organism' (Fig. 15).

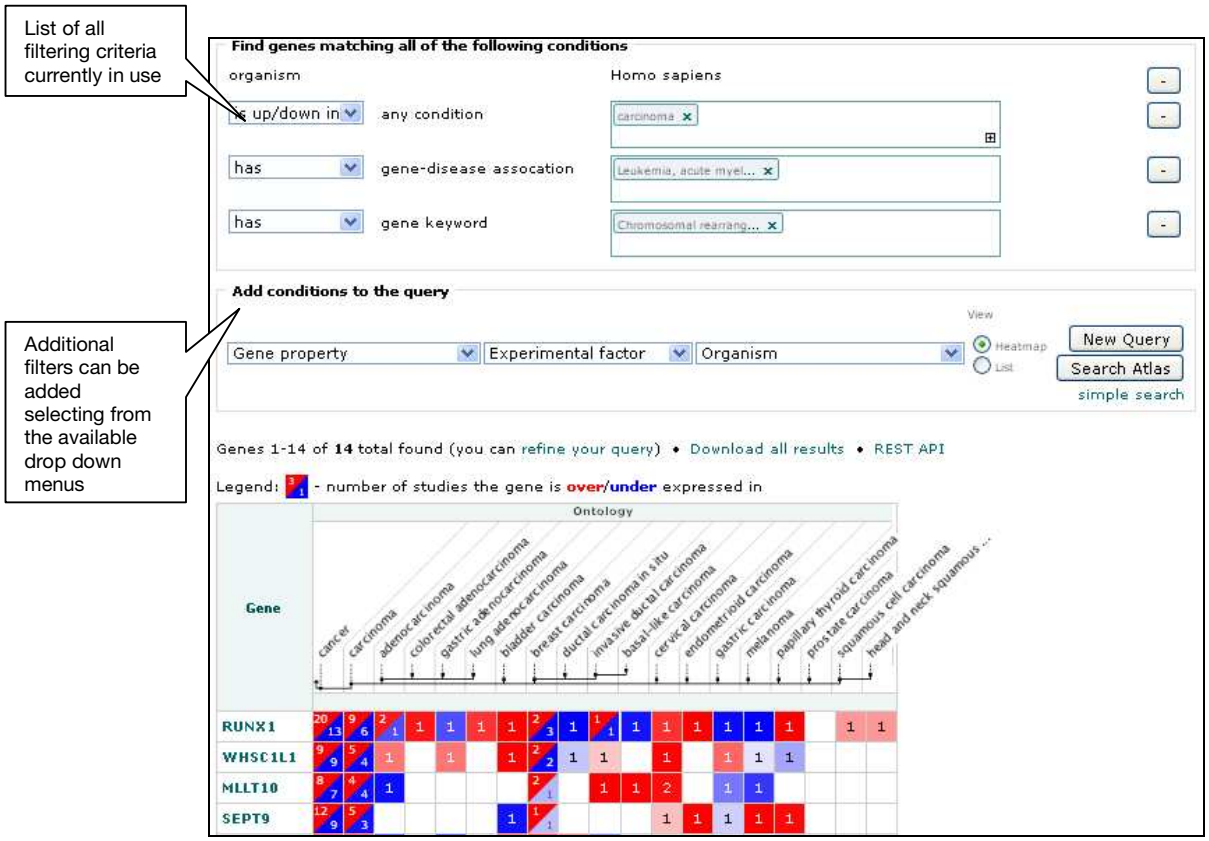


Fig. 15: Example of a condition query using the term 'carcinoma, EFO_0000313; ***additional filters on the terms 'Leukemia, acute myeloid' and 'Chromosomal rearrangement' have been added using the 'Refine your query' option.***

5 The AE Atlas advance search

We will now use the Atlas advance search to perform a more complex query.

1. Open the Atlas homepage, at http://www.ebi.ac.uk/microarray-as/atlas/, in a Web browser
2. Click on the 'advance search' link, on the right hand side of the page.

Three drop-down menus are available: 'Gene property', 'Experimental factor' and 'Organism' and they can be used to select your searching criteria (Fig. 16).

Fig. 16: Atlas advance search interface.

Let us search for genes which are up-regulated in 'hepatocellular carcinoma' AND down-regulated in 'liver' AND down-regulated in 'normal' (Fig. 17)

3. From the drop-down menu 'Experimental factor', select 'disease state'. The selected filter will now appear above, under 'Find genes matching all the following conditions'. Restrict the search to the gene(s) which 'is up in' and type 'hepatocellular carcinoma' in the text box on the right (auto-complete function available).
4. From the drop-down menu 'Experimental factor', select 'organism part'. This time restrict the search to the gene(s) which are 'down in' and type 'liver' in the text box.
5. Finally, from the dropdown menu 'Experimental factor', select again 'disease state' and restrict the search to the gene(s) which 'is down in' and type 'normal' in the text box.
6. Select the heatmap view and click on 'Search Atlas'.

A heatmap of all genes which match these searching criteria is retrieved and can be browsed as previously explained (Fig. 17).

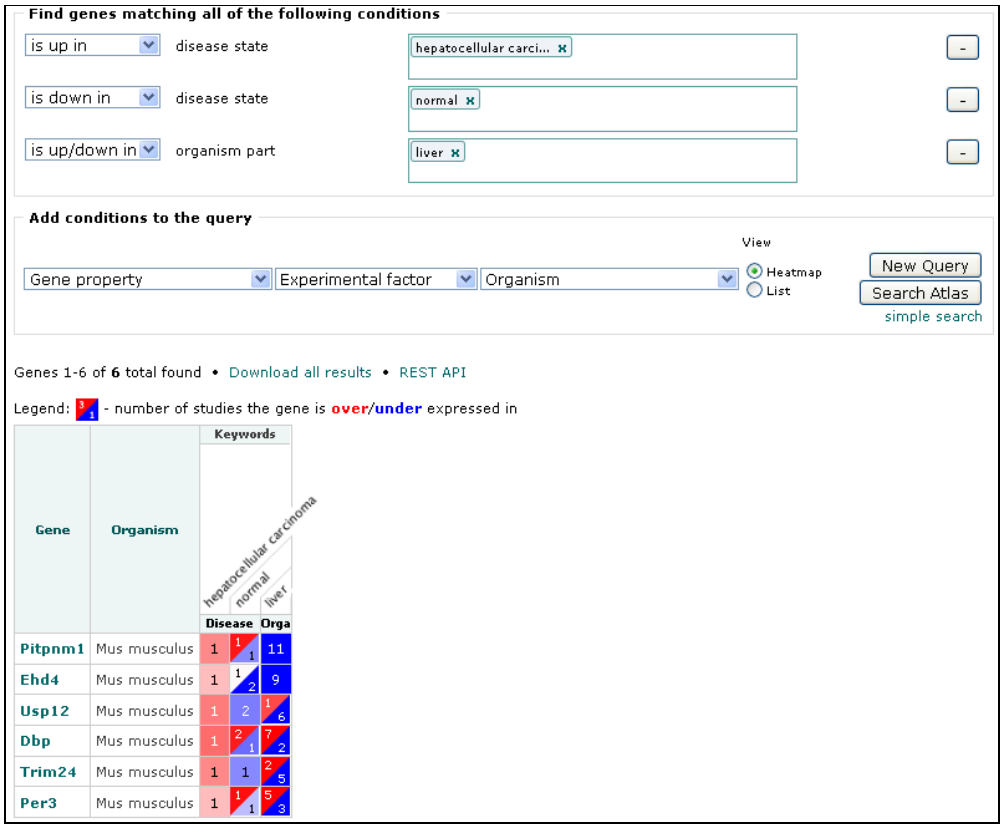


Fig.17: Atlas advance search output.

Glossary

Aspm

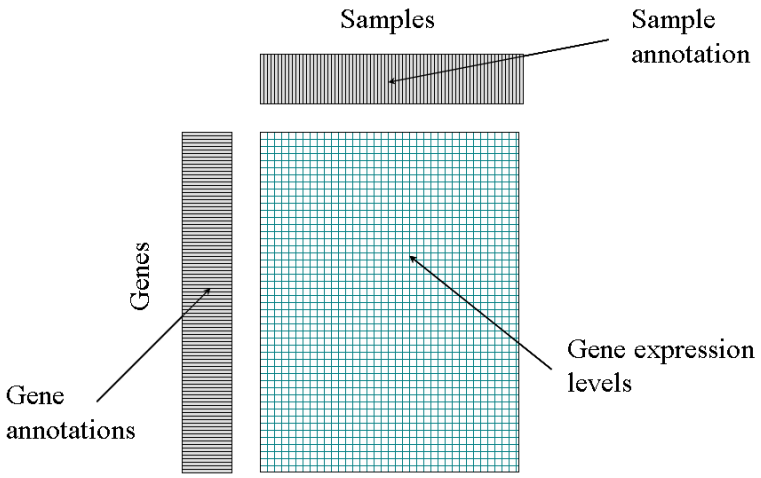
Abnormal spindle-like microcephaly-associated protein homolog. More information about this gene can be found in Ensembl at http://www.ensembl.org/Mus_musculus/Gene/Summary?g=ENSMUSG00000033952

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.



Differentially expressed

A gene is **differentially expressed** when its expression values under two or more conditions are statistically significantly different.

EFO

EFO stands for **Experimental Factor** Ontology and it is s an application focused ontology modelling the **experimental factor**s in AE (http://www.ebi.ac.uk/microarray-srv/efo).

Experiment

The complete set of hybridizations performed in a study.

Experimental factor or condition

A property that varies between samples and it is important in the interpretation of your data (e.g. time, compound, genotype, etc.).

Expression profile

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

Gene Annotation

Gene annotation is the process of attaching biological information to genes.

HUGO names

The HUGO gene nomenclature committee approves a gene name and symbol (short-form abbreviation) for each known human gene. All approved symbols are stored in the HGNC database (http://www.genenames.org/). Each symbol is unique and each gene is only given one approved gene symbol.

Meta-analytical

Defines a statistical method that combines the results of a number of different studies in order to provide a larger sample size for evaluation and to produce a stronger conclusion than can be provided by any single study

Ortholog

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution.

p-value

The probability of an event or outcome in a statistical experiment

Saa4

Serum amyloid A-4 protein precursor. More information about this gene can be found in Ensembl at http://www.ensembl.org/Mus_musculus/geneview?gene=ENSMUSG00000040017.

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., *ArrayExpress--a public repository for microarray **gene expression** data at the EBI*. Nucleic Acids Res, 2003. **31**(1): p. 68-71.
2. 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.