

- Helen Parkinson, Anna Farne, Misha Kapushesky (version1.0, 23/07/2008)

ArrayExpress Atlas for beginners

The ArrayExpress (AE) database is a public curated repository of gene expression and other array data, such as comparative genome hybridization data and Chip-on-chip data. The ArrayExpress Atlas uses ArrayExpress data and ontologies to present gene level queries and ranking of gene variability across experiments in the ArrayExpress data warehouse. 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 ArrayExpress atlas and how to use it
- 2 How to query the Atlas by genes
- 3 How to query the Atlas by conditions

1 What is the ArrayExpress Atlas and how to use it?

The Atlas provides a way to integrate gene expression data from GEO and ArrayExpress and provide gene and condition queries for these data. The Atlas uses data from the ArrayExpress data warehouse which have been carefully re-annotated for comparability and have had their gene annotation updated using [Uniprot](#) or [Ensembl](#). This means that we have added information such as [GO terms](#), [HGHC gene names](#) and other functional information where available. We are building an ontology of experimental variables, called the [EFO](#), and using it to structure the annotations for the data in the Atlas. This ontology is used to enhance Atlas queries. It can from the Ontology Look-Up service. The EFO is also a new resource and we welcome feedback, see <http://www.ebi.ac.uk/microarray-srv/efo> for details.

For every experiment in the ArrayExpress warehouse the strength of differential expression of genes across conditions across experiments is calculated and meta analytical statistics across experiments are calculated. The resulting data are available via a new user interface (June 2008). The Atlas is currently a beta release and we welcome your comments and questions so we can improve it.

2 How to query the ArrayExpress Atlas by Gene

The Atlas can be queried with gene attributes or experimental conditions (or variables) and provides a ranked list of conditions and experiments in which they are tested.



To familiarise you with the query form, we will perform a simple search querying first for a gene

1. Open <http://www.ebi.ac.uk/microarray-as/atlas/>
2. In the genes box type MAT1A – you can find out about this [gene here](#) from the Ensembl record
3. Select the default up/down and choose Homo sapiens
4. Leave the conditions box empty for now

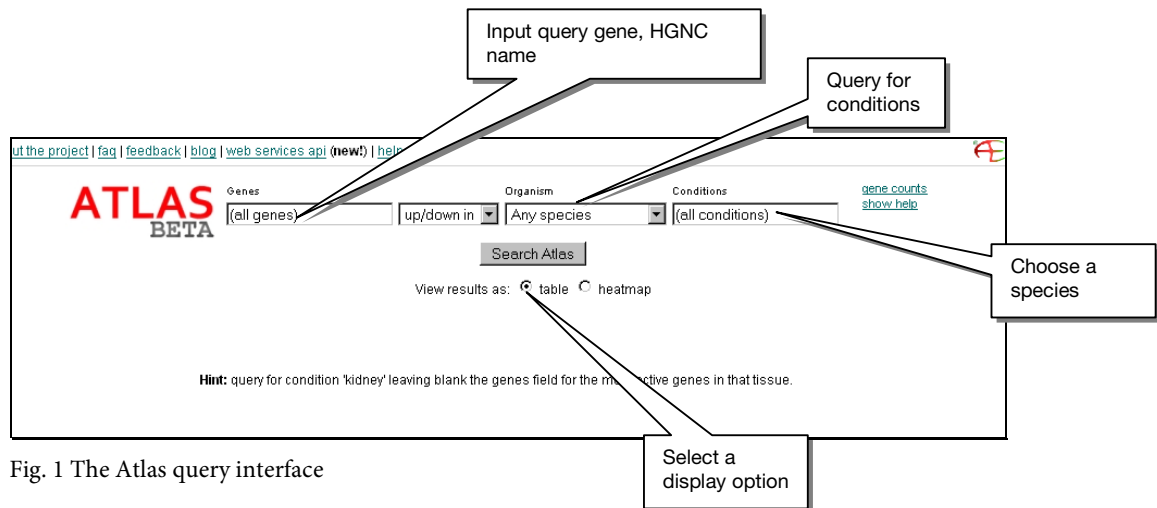


Fig. 1 The Atlas query interface

The screenshot shows the ATLAS BETA query result for MAT1A in Homo sapiens. Callouts point to the following features:

- List of experiments:** Points to the table header.
- Link to experiment Summary:** Points to the 'More...' link in the first row.
- Significantly over-expressed:** Points to the red background and upward arrow in the 'P-value' column for the first two rows.
- Significantly under-expressed:** Points to the blue background and downward arrow in the 'P-value' column for the second row.
- Link to expression profile for this gene for this experiment in Data warehouse:** Points to the 'More...' link in the first row.
- Search across all EBI:** Points to the 'More...' link in the last row.
- Synonyms searched:** Points to the 'gene_synonym' column.

Experiment	Description	Factor Value (Factor)	Gene Name	Gene Id	Organism	P-value	AEW	...
E-AFMX-5	Transcription profiling of human cell lines and tissues (GNF/Novartis)	liver (organismpart)	MAT1A	ENSG00000151224	Homo sapiens	↑ < 1e-16	More...	gene_name: MAT1A gene_synonym: MAT1A;AMS1;MATA1
E-AFMX-5	Transcription profiling of human cell lines and tissues (GNF/Novartis)	white blood cells (organismpart)	MAT1A	ENSG00000151224	Homo sapiens	↓ < 1e-16	More...	gene_name: MAT1A gene_synonym: MAT1A;AMS1;MATA1
E-GEOD-412	Transcription profiling of human leukemia cells treated with mercaptopurine and/or methotrexate	none (compound)	MAT1A	ENSG00000151224	Homo sapiens	↑ 4.08e-10	More...	gene_name: MAT1A gene_synonym: MAT1A;AMS1;MATA1
E-GEOD-3189	Transcription profiling of human tissue samples to identify novel genes associated with malignant melanoma but not benign melanocytic lesions	malignant melanoma (diseasestate)	MAT1A	ENSG00000151224	Homo sapiens	↑ 4.70e-09	More...	gene_name: MAT1A gene_synonym: MAT1A;AMS1;MATA1
E-GEOD-3254	Transcription profiling of Stratagene human universal reference RNA labelled with different labelling protocol_types to test labelling protocol_types	Enzo Labeling (protocoltype)	MAT1A	ENSG00000151224	Homo sapiens	↑ 6.24e-09	More...	gene_name: MAT1A gene_synonym: MAT1A;AMS1;MATA1

Figure 2. Example of an Atlas query result

The interface returns the list of all experiments (studies) in Atlas where the selected gene is up or down regulated in any of the conditions, in this case in human (Fig. 3). All the conditions are ordered by 'relevance', with the most relevant experiment on top. The 'relevance rank' is based on the correlation between experimental factors values and gene expression values and it is calculated via a

linear model in the Bioconductor package LIMMA [Smyth 2004]. For each condition, a short description, experimental factors and p-value are provided. Blue indicates relative under-expression and red indicates over-expression.

Questions:

1. What types of experiment is variable MAT1A expression detected in?
2. What tissue(s) is MAT1A expressed in?
3. What are the likely functions of MAT1A?
4. What other genes share a similar expression profile with MAT1A in Experiment E-AFMX-5 – click ‘more’ and do a similarity search
5. If you query the Atlas for these genes are they expressed in the tissue types the same as for MAT1A?

3. How to query the ArrayExpress Atlas by condition

The condition box allows auto-complete and also queries using an ontology to get to child terms, e.g. breast cancer –is-a (kind of) cancer. Standard queries use string matching using Lucene technology.

1. Query conditions with ‘cancer’

The screenshot shows the ATLAS BETA search interface. The search bar contains 'cancer' under the 'Conditions' field. Below the search bar, a list of expanded child terms is shown, including various cancer types like 'cancer', 'monophasic synovial sarcoma', 'chondrosarcoma', etc. Below the list, a table displays search results for two experiments: E-GEOD-1872 and E-GEOD-2712. The table columns are Experiment, Description, Factor Value (Factor), Gene Name, Gene Id, Organism, P-value, and AEW. The P-value for both experiments is highlighted in blue as < 1e-16.

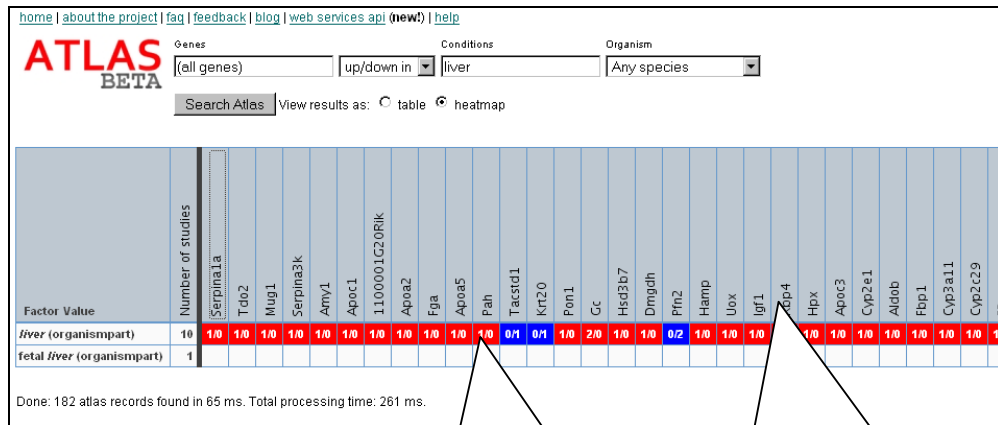
Experiment	Description	Factor Value (Factor)	Gene Name	Gene Id	Organism	P-value	AEW
E-GEOD-1872	Transcription profiling of rat mammary tumors induced by N-methyl-N-nitrosourea (NMU) and normal mammary gland	breast cancer (diseasestate)	Agt	ENSRNOG00000018445	Rattus norvegicus	< 1e-16	More...
E-GEOD-2712	Transcription profiling of human Clear cell sarcoma of the kidney (CCSK) vs Wilms tumors	clear cell sarcoma of the kidney	WT1	ENSG00000184937	Homo sapiens	< 1e-16	More...

Figure 3. Example of a condition query using ‘cancer’

2. What are the most common conditions retrieved by this query
3. Try a tissue specific query for ‘liver’ – which genes are returned?
4. For queries returning a lot of genes try the heat map view, click heat-map and ‘search’

Query has been expanded to all child terms present in ArrayExpress data annotated with EFO

5. How many genes are returned as over-expressed in human liver?



Number of experiments score, red is over-expressed, blue is under expressed

Links back to the ArrayExpress warehouse

Further reading

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29(4):365-71
2. Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3(12):Article3.