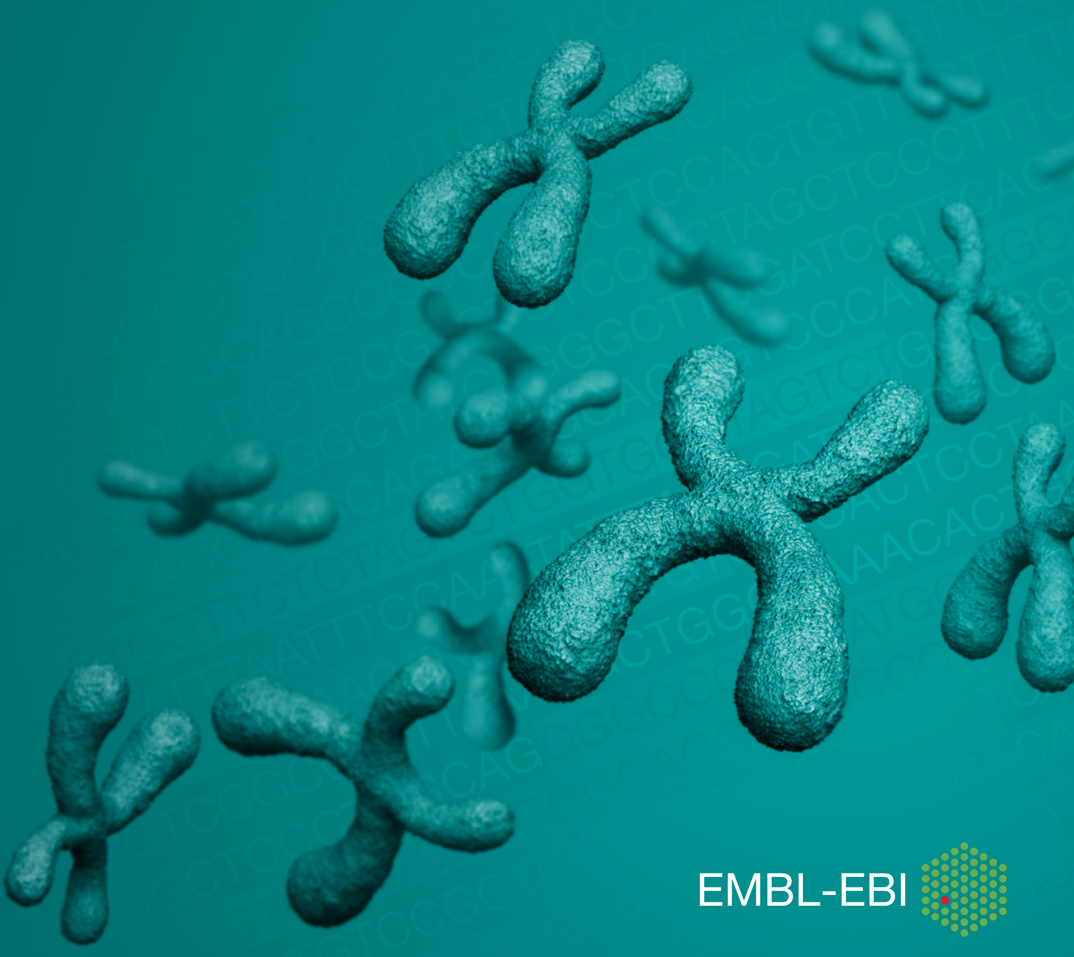
ArrayExpress and Expression Atlas: Mining Functional Genomics data

Gabriella Rustici, PhD

Functional Genomics Team
EBI-EMBL

gabry@ebi.ac.uk



EMBL-EBI

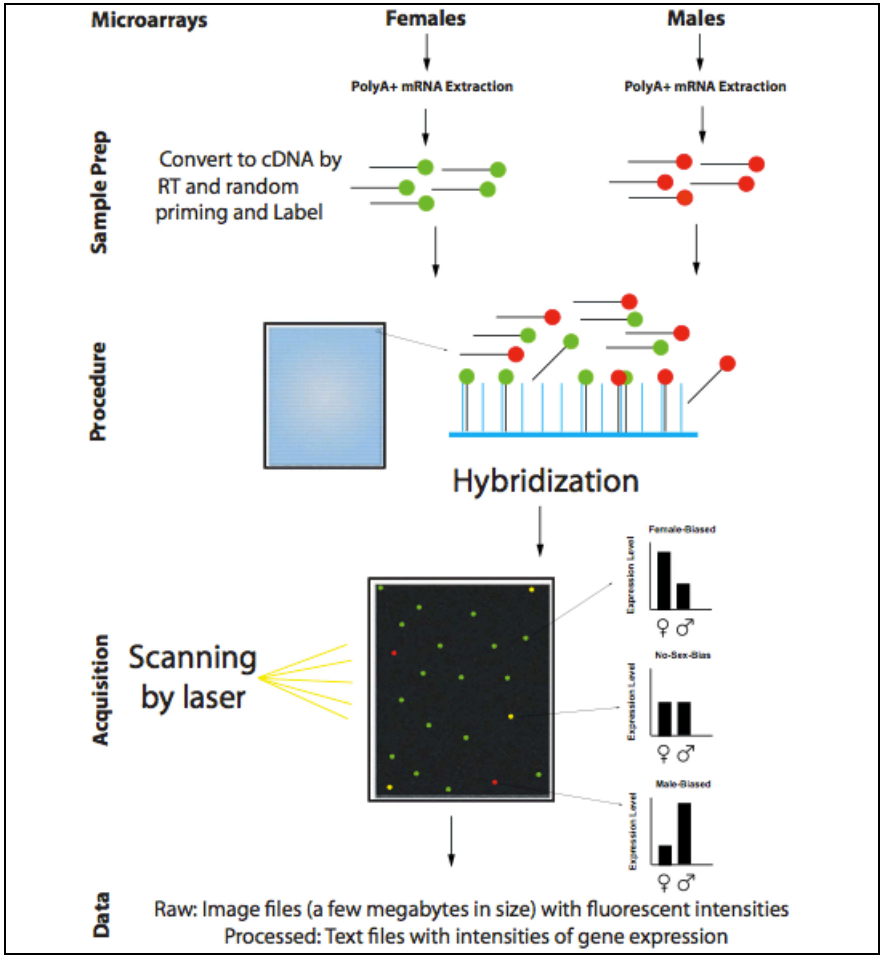
What is functional genomics (FG)?

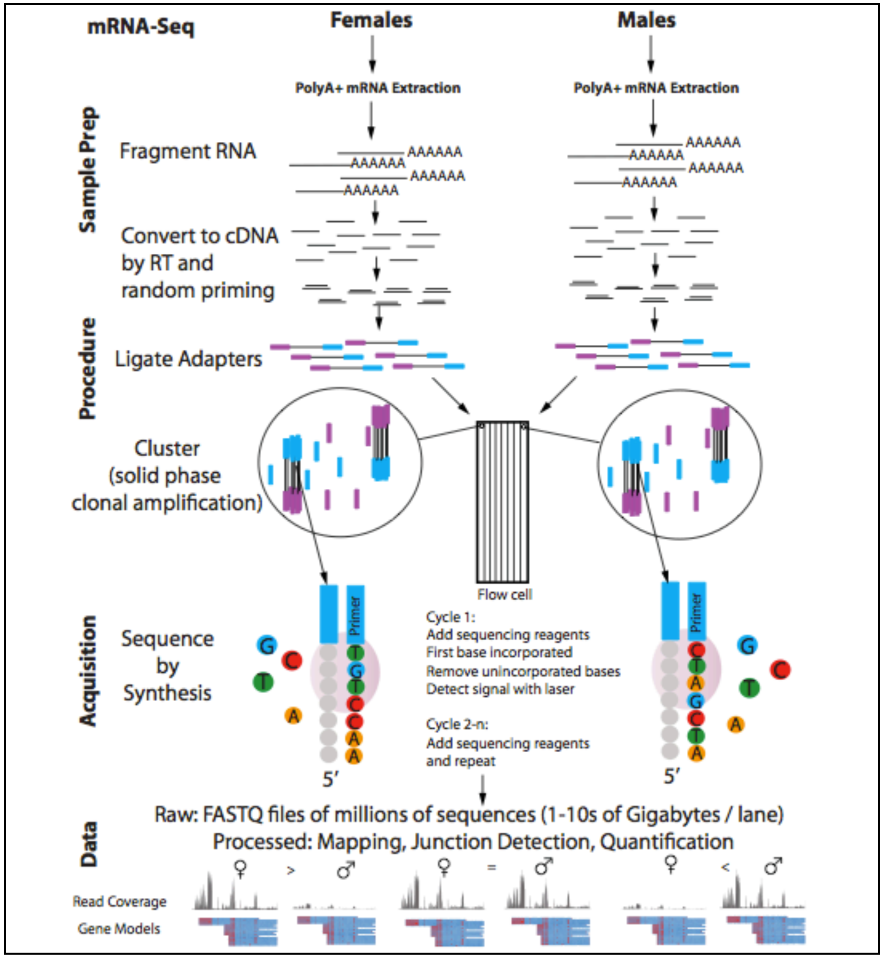
- The aim of FG is to understand the function of genes and other parts of the genome
- FG experiments typically utilize genome-wide assays to measure and track many genes (or proteins) in parallel under different conditions
- High-throughput technologies such as microarrays and high-throughput sequencing (HTS) are frequently used in this field to interrogate the transcriptome

What biological questions is FG addressing?

- When and where are genes expressed?
- How do gene expression levels differ in various cell types and states?
- What are the functional roles of different genes and in what cellular processes do they participate?
- How are genes regulated?
- How do genes and gene products interact?
- How is gene expression changed in various diseases or following a treatment?

Components of a FG experiment





ArrayExpress
www.ebi.ac.uk/arrayexpress/

- Is a public repository for FG data, which provides easy access to well annotated data in a structured and standardized format
- Serves the scientific community as an archive for data supporting publications, together with GEO at NCBI and CIBEX at DDBJ
- Facilitates the sharing of experimental information associated with the data such as microarray designs, experimental protocols,……
- Based on community standards: MIAME guidelines & MAGE-TAB format for microarray, MINSEQE guidelines for HTS data (http://www.fged.org/minseqe/)

Community standards for data requirement

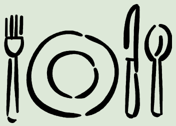
- MIAME = Minimal Information About a Microarray Experiment
- MINSEQE = Minimal Information about a high-throughput Nucleotide SEQuencing Experiment
- The checklist:

Requirements	MIAME	MINSEQE
1. Experiment design / background description	✓	✓
2. Sample annotation and experimental factor	✓	✓
3. Array design annotation (e.g. probe sequence)	✓	
4. All protocols (wet-lab bench and data processing)	✓	✓
5. Raw data files (from scanner or sequencing machine)	✓	✓
6. Processed data files (normalised and/or transformed)	✓	✓

What is an experimental factor?

- The main variable(s) studied, often related to the hypothesis of the experiment and is the independent variable.

- Values of the factor ("factor values") should vary.

Experimental design	Factor ✓	Factor Values	Not factor ×
 beef vs horse meat	 Diet	beef, horse meat	Organism (human)
 smoker vs non-smoker	 compound	cigarette smoke (tobacco), no tobacco	Organism (human), sex (male)
 face cream A vs control X	compound	Active ingredient A, "sham" control	Cell type

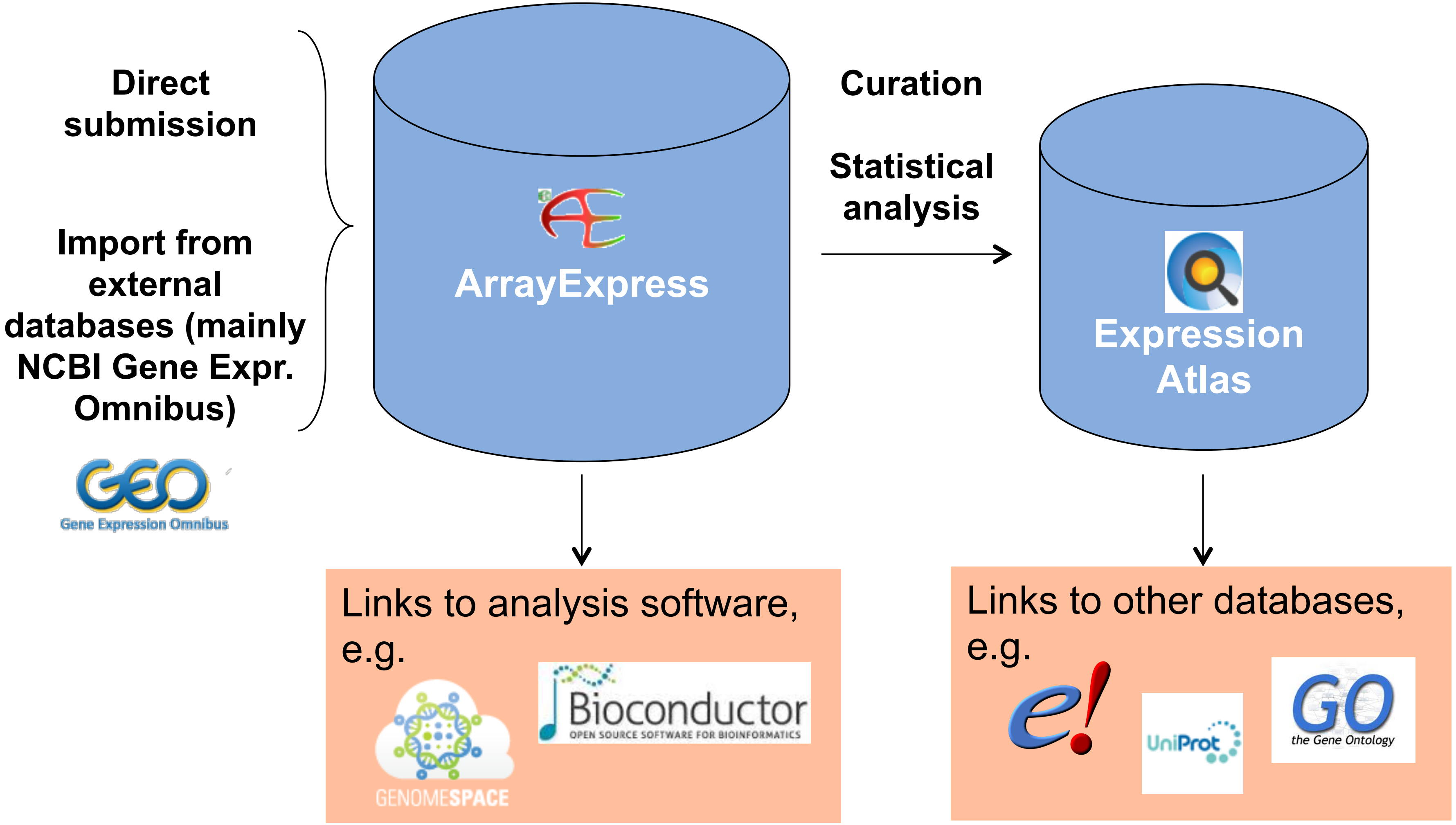
Standards for microarray & sequencing

MAGE-TAB format

MAGE-TAB is a simple spreadsheet format that uses a number of different files to capture information about a microarray or sequencing experiments

IDF	**Investigation Description Format** file, contains top-level information about the experiment including title, description, submitter contact details and protocols.
SDRF	**Sample and Data Relationship Format** file contains the relationships between samples and arrays, as well as sample properties and experimental factors, as provided by the data submitter.
ADF	**Array Design Format** file that describes probes on an array, e.g. sequence, genomic mapping location (for array data ony)
Data files	**Raw and processed data** files. The 'raw' data files are the trace data files (.srf or .sff). Fastq format files are also accepted, but SRF format files are preferred. The trace data files that you submit to ArrayExpress will be stored in the European Nucleotide Archive (ENA). The processed data file is a 'data matrix' file containing processed values, e.g. files in which the expression values are linked to genome coordinates.

The two databases: how are they related?



Direct submission

Import from external databases (mainly NCBI Gene Expr. Omnibus)

ArrayExpress

Curation

Statistical analysis

Expression Atlas

Links to analysis software, e.g.

Links to other databases, e.g.

What is the difference between them?

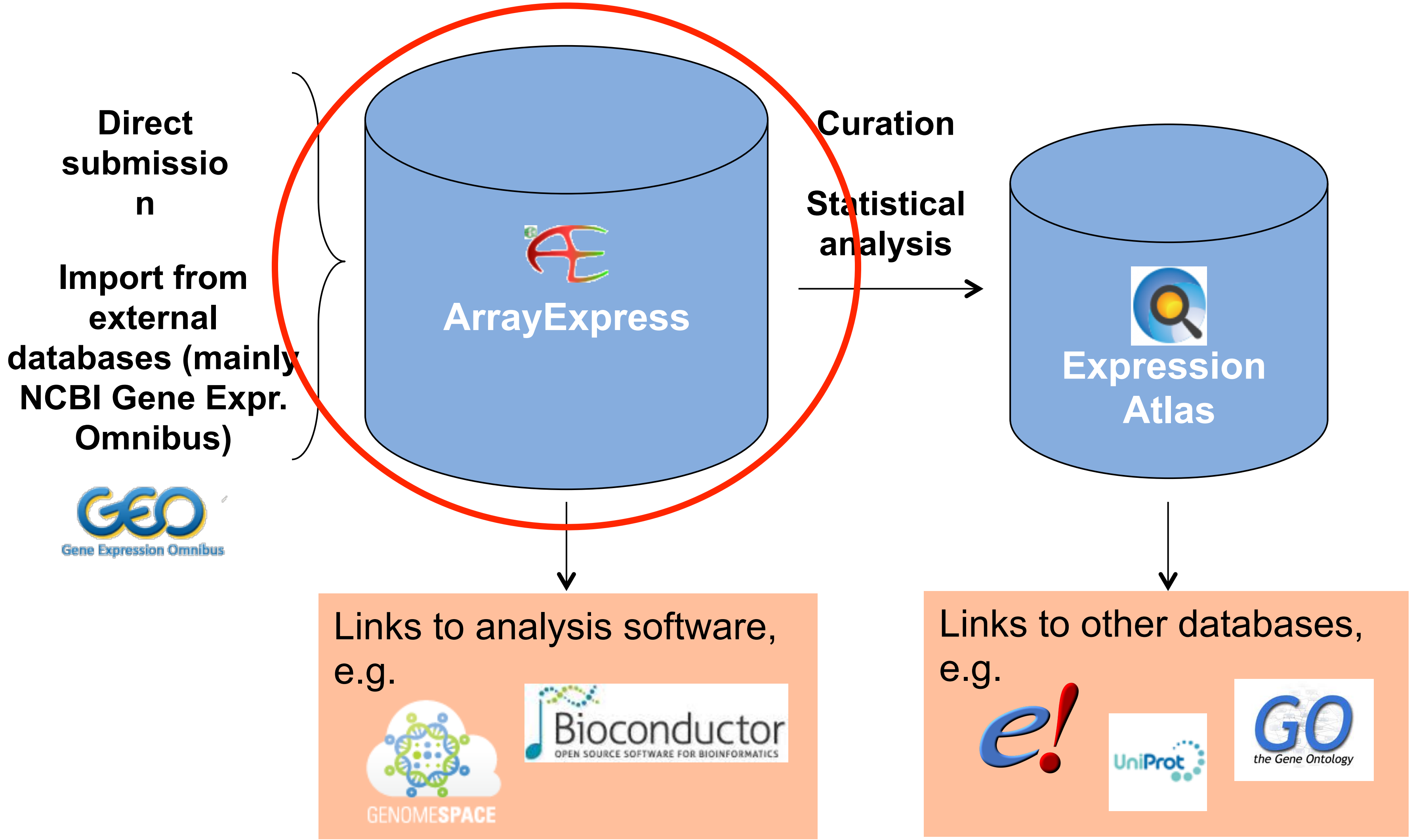
ArrayExpress Archive

- Central object: experiment
- Query to retrieve experimental information and associated data

Expression Atlas

- Central object: gene/condition
- Query for gene expression changes across experiments and across platforms

The two databases: how are they related?



Direct submission

Import from external databases (mainly NCBI Gene Expr. Omnibus)

Curation

Statistical analysis

Links to analysis software, e.g.

Links to other databases, e.g.

ArrayExpress Archive – when to use it?

- Find FG experiments that might be relevant to your research

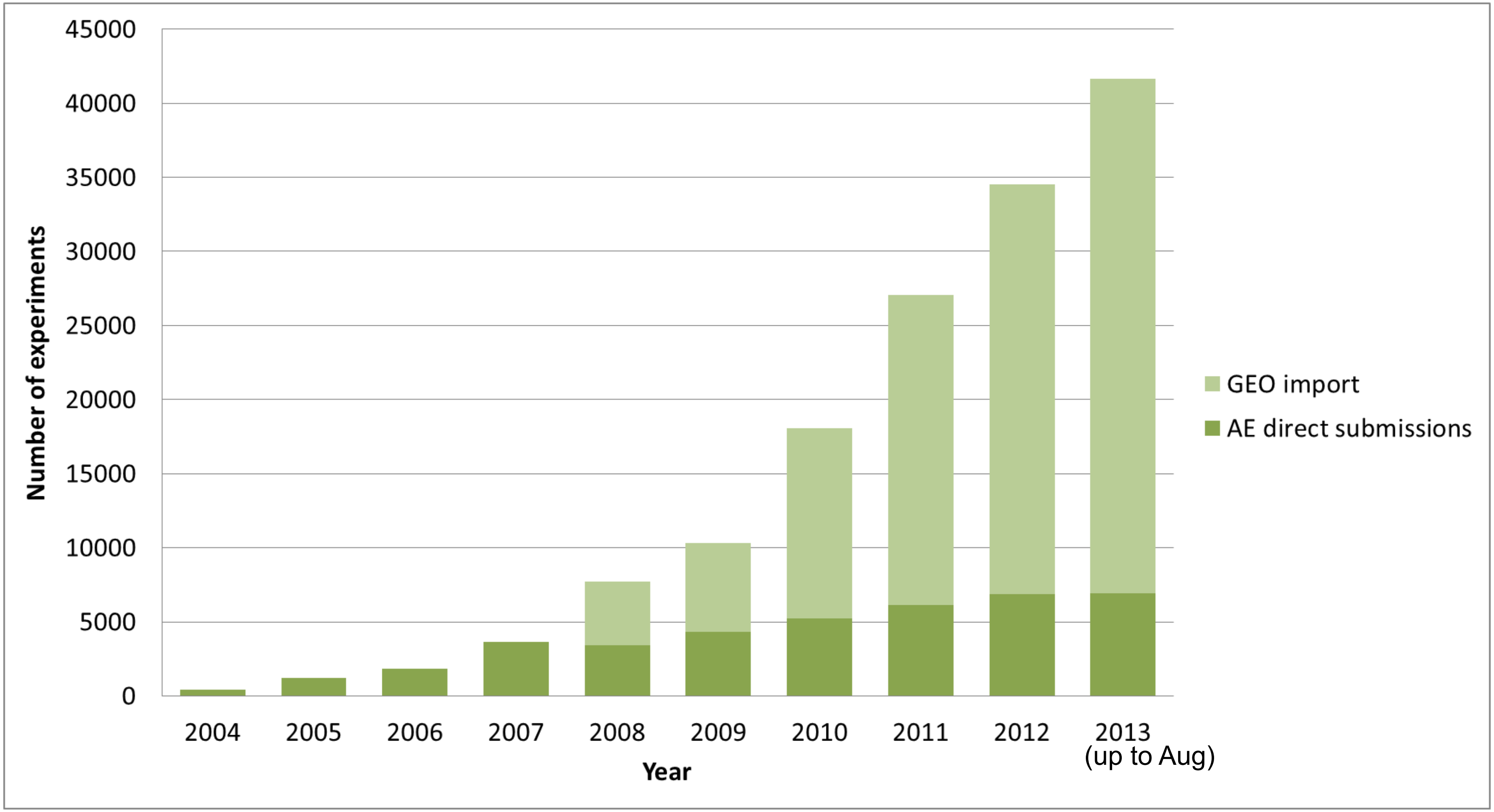
- Download data and re-analyze it.

  Often data deposited in public repositories can be used to answer different biological questions from the one asked in the original experiments.

- Submit microarray or HTS data that you want to publish.

  Major journals will require data to be submitted to a public repository like ArrayExpress as part of the peer-review process.

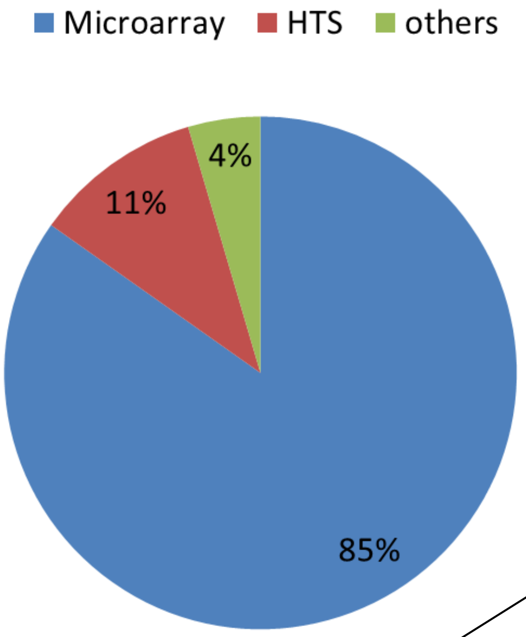
How much data in ArrayExpress?
(as of late August 2013)



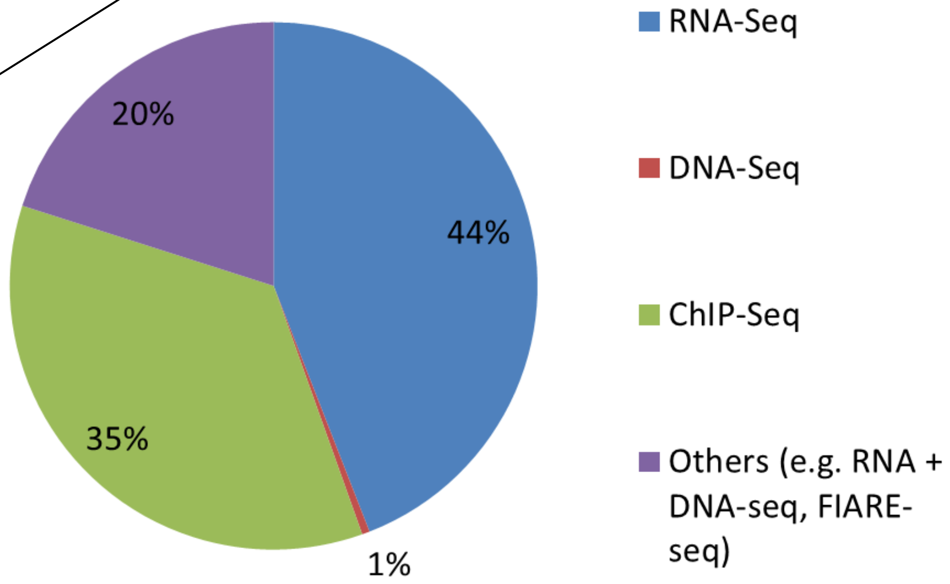
Breakdown of microarray vs seq. data

(as of late August 2013)

Microarray vs HTS

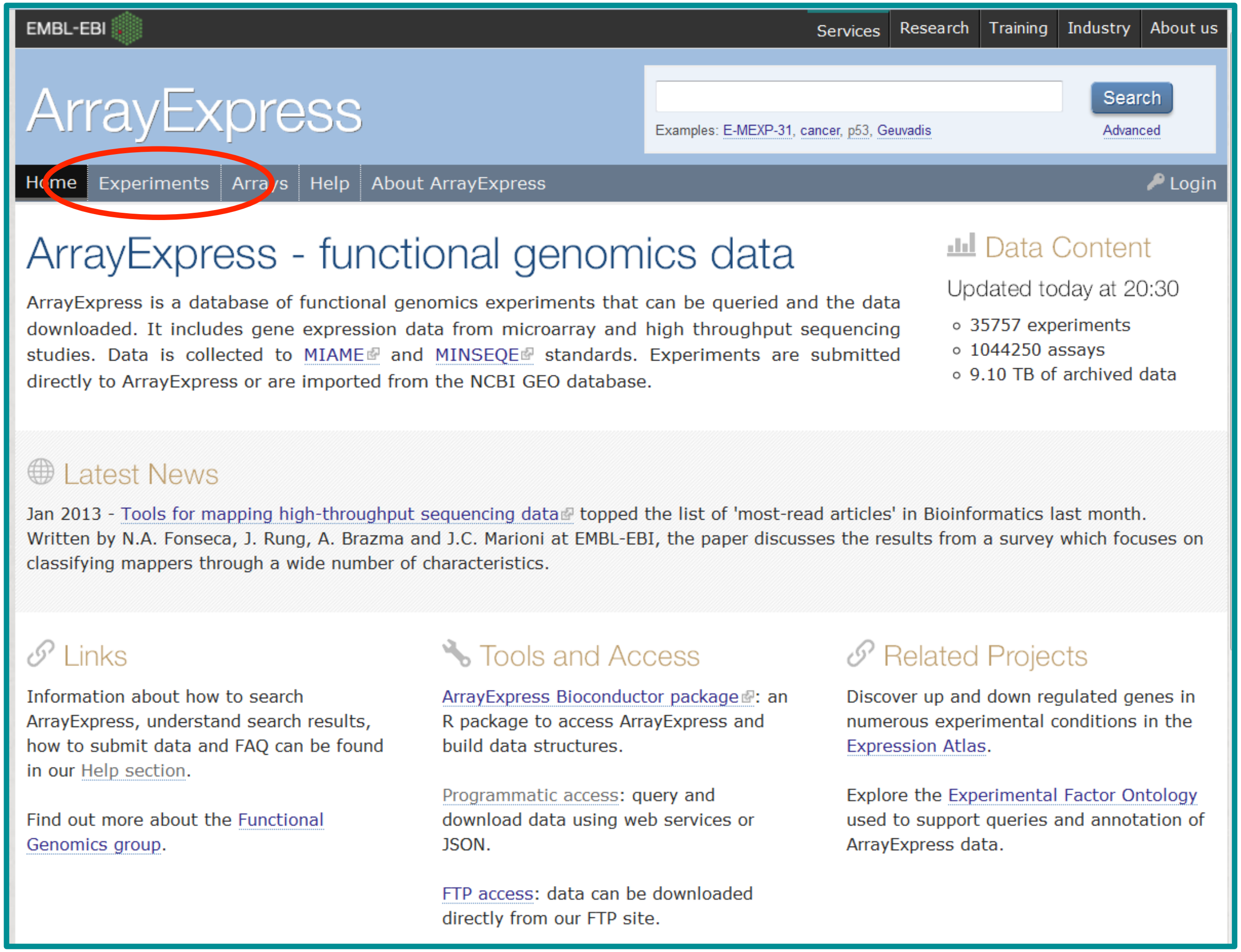


RNA-, DNA-, ChIP-seq breakdown



Browsing ArrayExpress

www.ebi.ac.uk/arrayexpress



Examples: E-MEXP-31, cancer, p53, Geuvadis

ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Data Content

Updated today at 20:30

- 35757 experiments
- 1044250 assays
- 9.10 TB of archived data

Latest News

Jan 2013 - Tools for mapping high-throughput sequencing data topped the list of 'most-read articles' in Bioinformatics last month. Written by N.A. Fonseca, J. Rung, A. Brazma and J.C. Marioni at EMBL-EBI, the paper discusses the results from a survey which focuses on classifying mappers through a wide number of characteristics.

Links

Information about how to search ArrayExpress, understand search results, how to submit data and FAQ can be found in our Help section.

Find out more about the Functional Genomics group.

Tools and Access

ArrayExpress Bioconductor package: an R package to access ArrayExpress and build data structures.

Programmatic access: query and download data using web services or JSON.

FTP access: data can be downloaded directly from our FTP site.

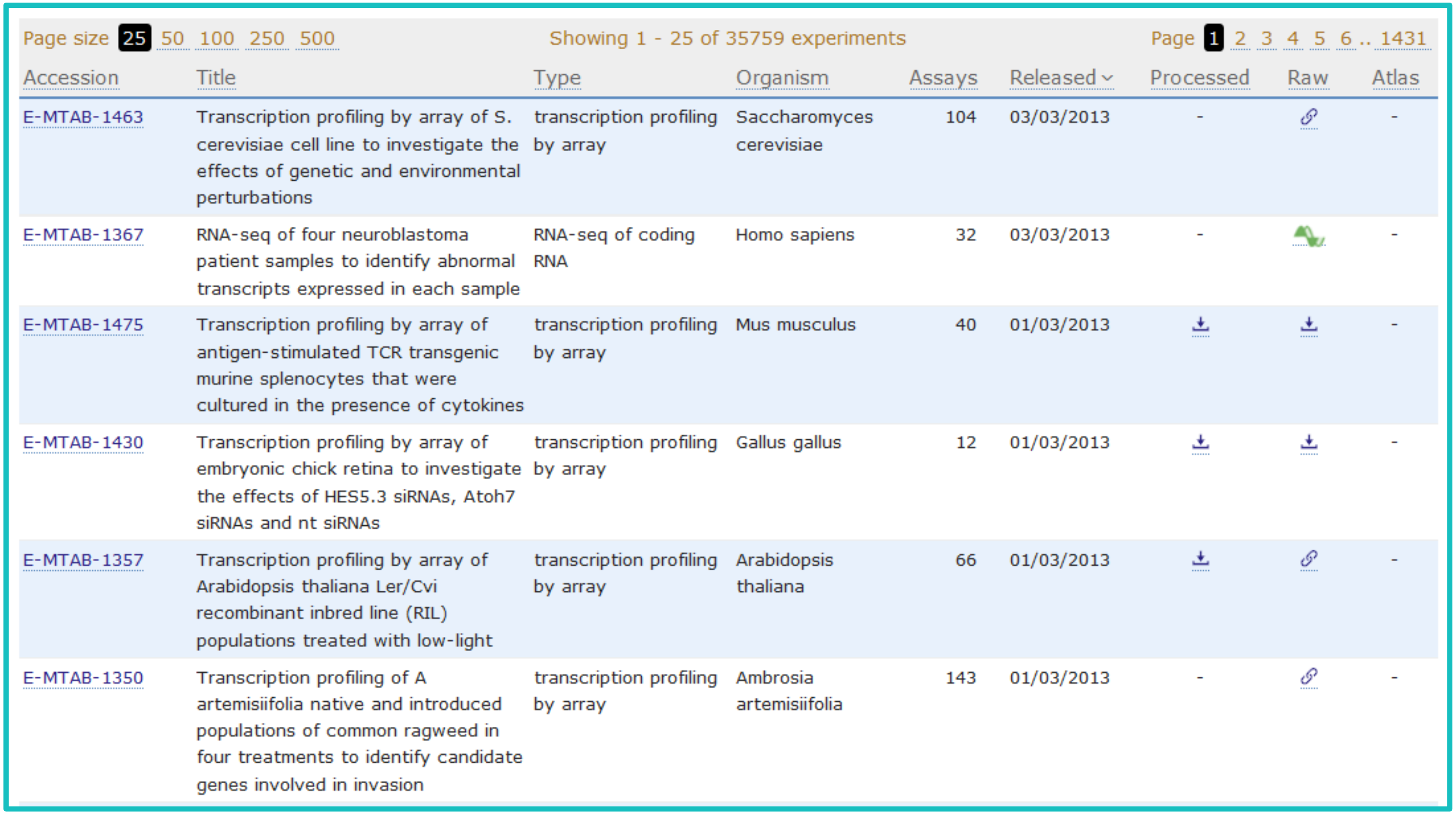
Related Projects

Discover up and down regulated genes in numerous experimental conditions in the Expression Atlas.

Explore the Experimental Factor Ontology used to support queries and annotation of ArrayExpress data.

Browsing ArrayExpress experiments

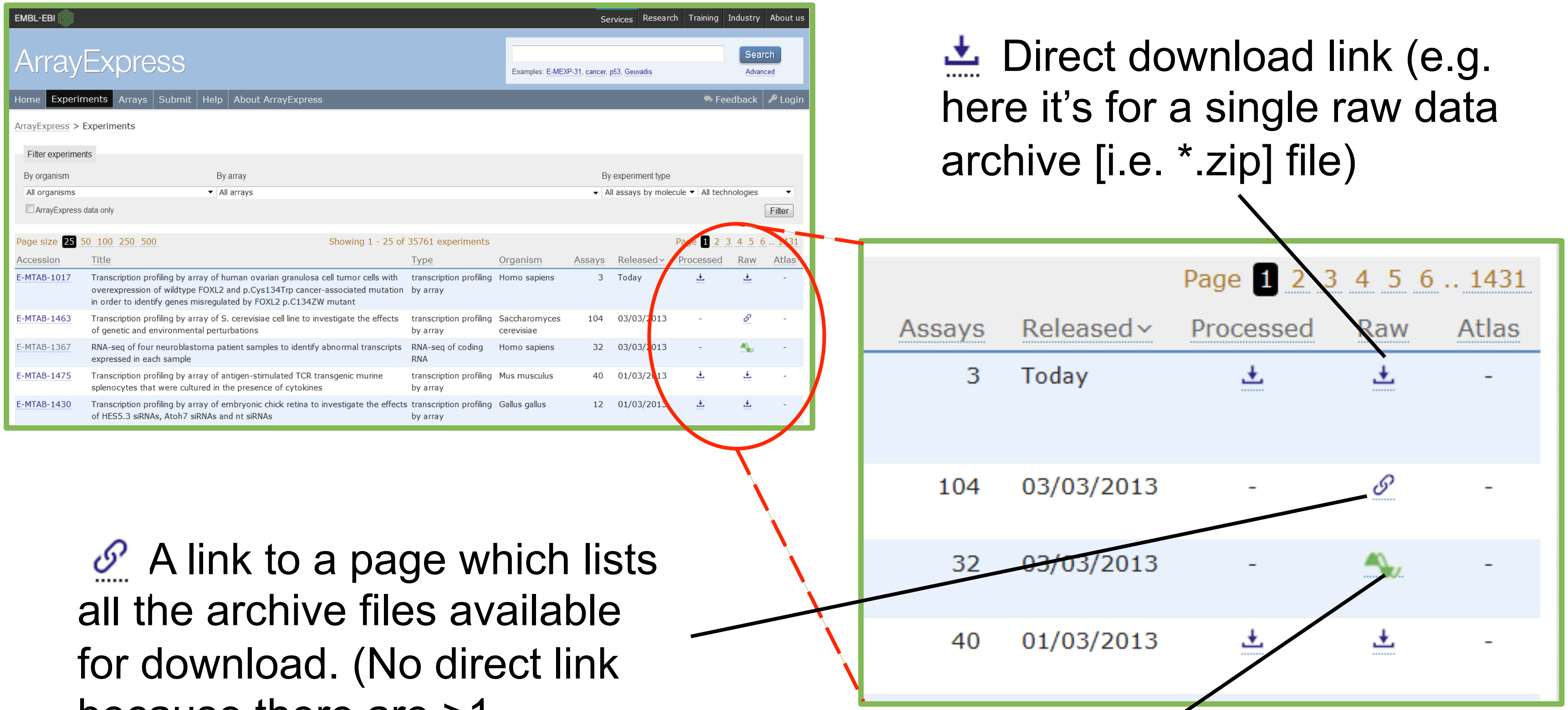
www.ebi.ac.uk/arrayexpress/experiments/browse.html



Page size 25 50 100 250 500 Showing 1 - 25 of 35759 experiments Page 1 2 3 4 5 6 .. 1431

Accession	Title	Type	Organism	Assays	Released ⌄	Processed	Raw	Atlas
E-MTAB-1463	Transcription profiling by array of S. cerevisiae cell line to investigate the effects of genetic and environmental perturbations	transcription profiling by array	Saccharomyces cerevisiae	104	03/03/2013	-	🔗	-
E-MTAB-1367	RNA-seq of four neuroblastoma patient samples to identify abnormal transcripts expressed in each sample	RNA-seq of coding RNA	Homo sapiens	32	03/03/2013	-	〰	-
E-MTAB-1475	Transcription profiling by array of antigen-stimulated TCR transgenic murine splenocytes that were cultured in the presence of cytokines	transcription profiling by array	Mus musculus	40	01/03/2013	⬇	⬇	-
E-MTAB-1430	Transcription profiling by array of embryonic chick retina to investigate the effects of HES5.3 siRNAs, Atoh7 siRNAs and nt siRNAs	transcription profiling by array	Gallus gallus	12	01/03/2013	⬇	⬇	-
E-MTAB-1357	Transcription profiling by array of Arabidopsis thaliana Ler/Cvi recombinant inbred line (RIL) populations treated with low-light	transcription profiling by array	Arabidopsis thaliana	66	01/03/2013	⬇	🔗	-
E-MTAB-1350	Transcription profiling of A artemisiifolia native and introduced populations of common ragweed in four treatments to identify candidate genes involved in invasion	transcription profiling by array	Ambrosia artemisiifolia	143	01/03/2013	-	🔗	-

File download on the Browse page



Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-MTAB-1017	Transcription profiling by array of human ovarian granulosa cell tumor cells with overexpression of wildtype FOXL2 and p.Cys134Trp cancer-associated mutation in order to identify genes misregulated by FOXL2 p.C134ZW mutant	transcription profiling by array	Homo sapiens	3	Today	⤓	⤓	-
E-MTAB-1463	Transcription profiling by array of S. cerevisiae cell line to investigate the effects of genetic and environmental perturbations	transcription profiling by array	Saccharomyces cerevisiae	104	03/03/2013	-	🔗	-
E-MTAB-1367	RNA-seq of four neuroblastoma patient samples to identify abnormal transcripts expressed in each sample	RNA-seq of coding RNA	Homo sapiens	32	03/03/2013	-	∿	-
E-MTAB-1475	Transcription profiling by array of antigen-stimulated TCR transgenic murine splenocytes that were cultured in the presence of cytokines	transcription profiling by array	Mus musculus	40	01/03/2013	⤓	⤓	-
E-MTAB-1430	Transcription profiling by array of embryonic chick retina to investigate the effects of HES5.3 siRNAs, Atoh7 siRNAs and nt siRNAs	transcription profiling by array	Gallus gallus	12	01/03/2013	⤓	⤓	-

Page 1 2 3 4 5 6 .. 1431

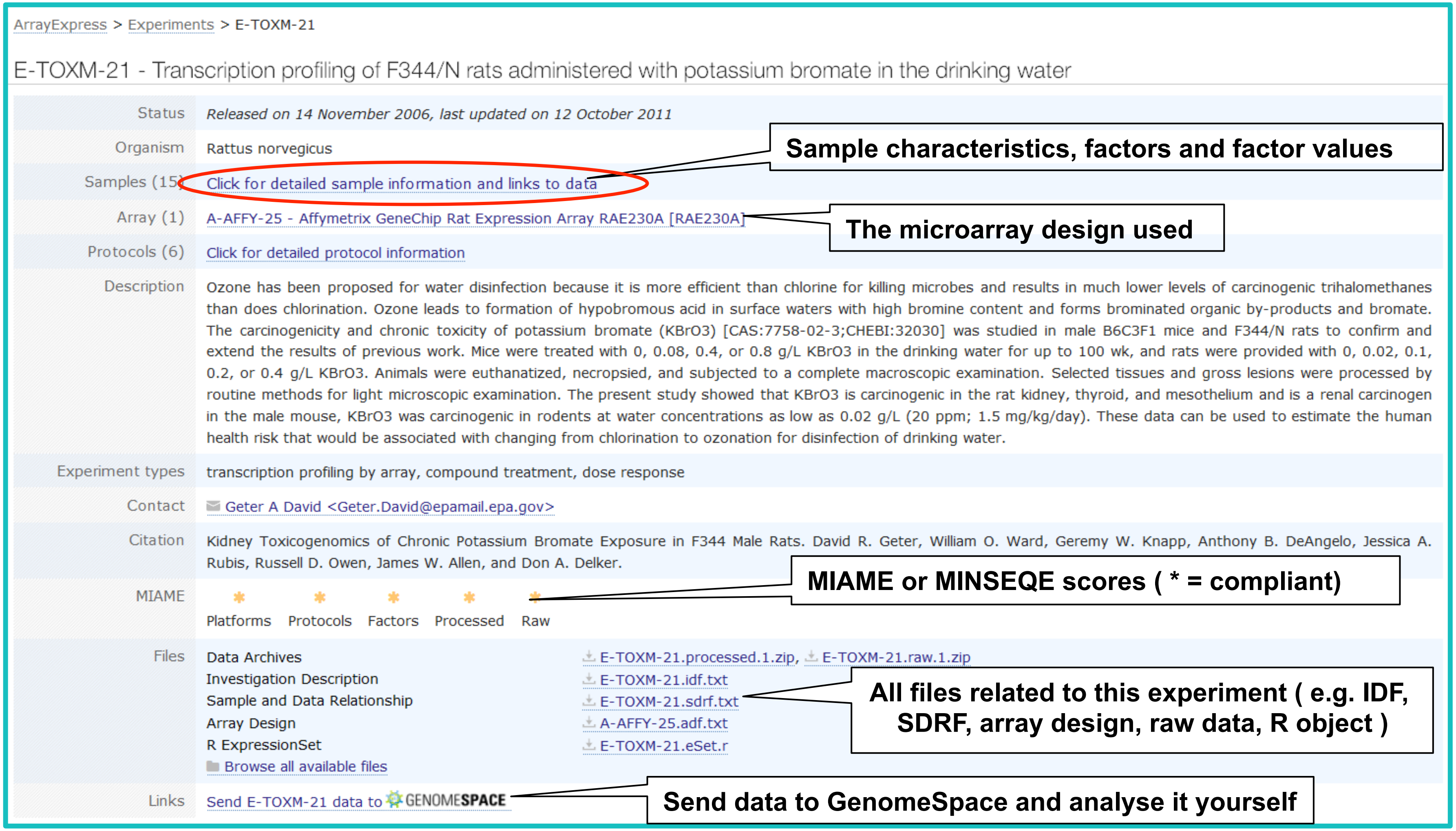
Assays	Released	Processed	Raw	Atlas
3	Today	⤓	⤓	-
104	03/03/2013	-	🔗	-
32	03/03/2013	-	∿	-
40	01/03/2013	⤓	⤓	-

⤓ Direct download link (e.g. here it's for a single raw data archive [i.e. *.zip] file)

🔗 A link to a page which lists all the archive files available for download. (No direct link because there are >1 archives)

∿ This is specifically for HTS experiments. Direct link to European Nucleotide Archive (ENA)'s page which lists all the sequencing assays (which are called "runs" at the ENA).

ArrayExpress single-experiment view



ArrayExpress > Experiments > E-TOXM-21

E-TOXM-21 - Transcription profiling of F344/N rats administered with potassium bromate in the drinking water

Status	*Released on 14 November 2006, last updated on 12 October 2011*
Organism	Rattus norvegicus
Samples (15)	Click for detailed sample information and links to data
Array (1)	A-AFFY-25 - Affymetrix GeneChip Rat Expression Array RAE230A [RAE230A]
Protocols (6)	Click for detailed protocol information
Description	Ozone has been proposed for water disinfection because it is more efficient than chlorine for killing microbes and results in much lower levels of carcinogenic trihalomethanes than does chlorination. Ozone leads to formation of hypobromous acid in surface waters with high bromine content and forms brominated organic by-products and bromate. The carcinogenicity and chronic toxicity of potassium bromate (KBrO3) [CAS:7758-02-3;CHEBI:32030] was studied in male B6C3F1 mice and F344/N rats to confirm and extend the results of previous work. Mice were treated with 0, 0.08, 0.4, or 0.8 g/L KBrO3 in the drinking water for up to 100 wk, and rats were provided with 0, 0.02, 0.1, 0.2, or 0.4 g/L KBrO3. Animals were euthanatized, necropsied, and subjected to a complete macroscopic examination. Selected tissues and gross lesions were processed by routine methods for light microscopic examination. The present study showed that KBrO3 is carcinogenic in the rat kidney, thyroid, and mesothelium and is a renal carcinogen in the male mouse, KBrO3 was carcinogenic in rodents at water concentrations as low as 0.02 g/L (20 ppm; 1.5 mg/kg/day). These data can be used to estimate the human health risk that would be associated with changing from chlorination to ozonation for disinfection of drinking water.
Experiment types	transcription profiling by array, compound treatment, dose response
Contact	✉ Geter A David <Geter.David@epamail.epa.gov>
Citation	Kidney Toxicogenomics of Chronic Potassium Bromate Exposure in F344 Male Rats. David R. Geter, William O. Ward, Geremy W. Knapp, Anthony B. DeAngelo, Jessica A. Rubis, Russell D. Owen, James W. Allen, and Don A. Delker.
MIAME	* Platforms * Protocols * Factors * Processed * Raw
Files	Data Archives: E-TOXM-21.processed.1.zip, E-TOXM-21.raw.1.zip Investigation Description: E-TOXM-21.idf.txt Sample and Data Relationship: E-TOXM-21.sdrf.txt Array Design: A-AFFY-25.adf.txt R ExpressionSet: E-TOXM-21.eSet.r Browse all available files
Links	Send E-TOXM-21 data to GENOMESPACE

Sample characteristics, factors and factor values

The microarray design used

MIAME or MINSEQE scores (* = compliant)

All files related to this experiment (e.g. IDF, SDRF, array design, raw data, R object)

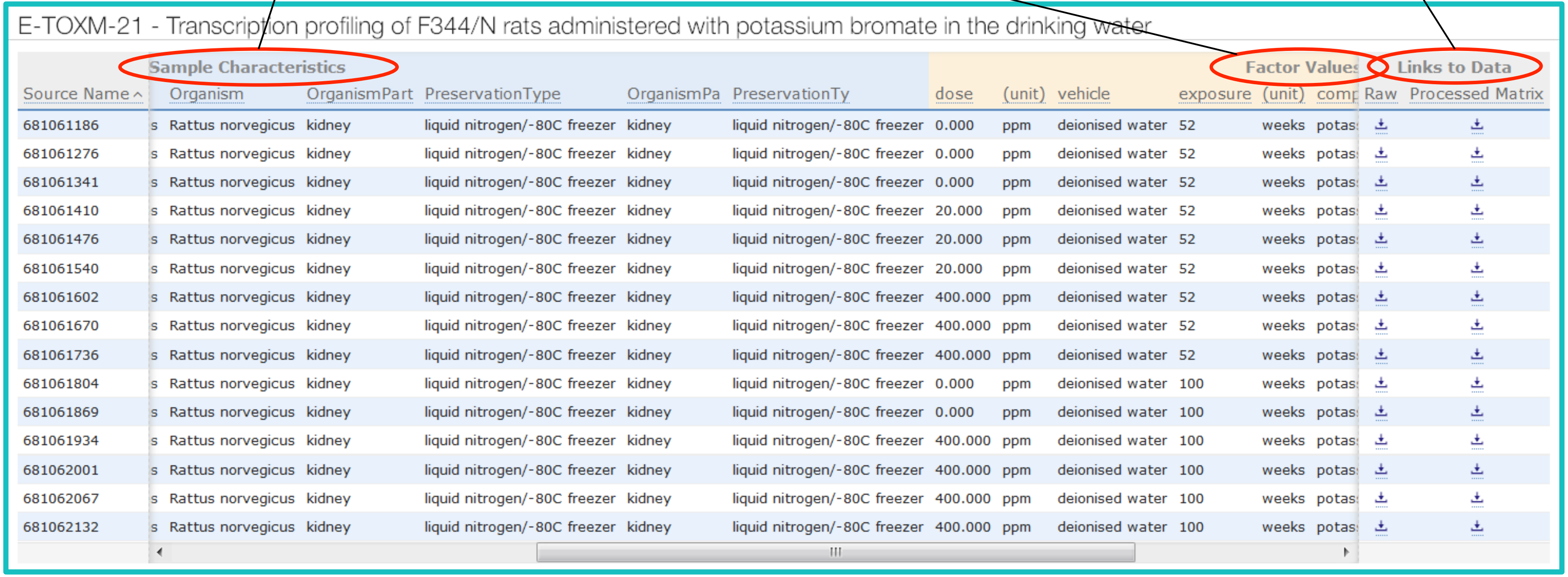
Send data to GenomeSpace and analyse it yourself

Samples view – microarray experiment

Sample characteristics

Factor values

Direct link to data files for one sample



E-TOXM-21 - Transcription profiling of F344/N rats administered with potassium bromate in the drinking water

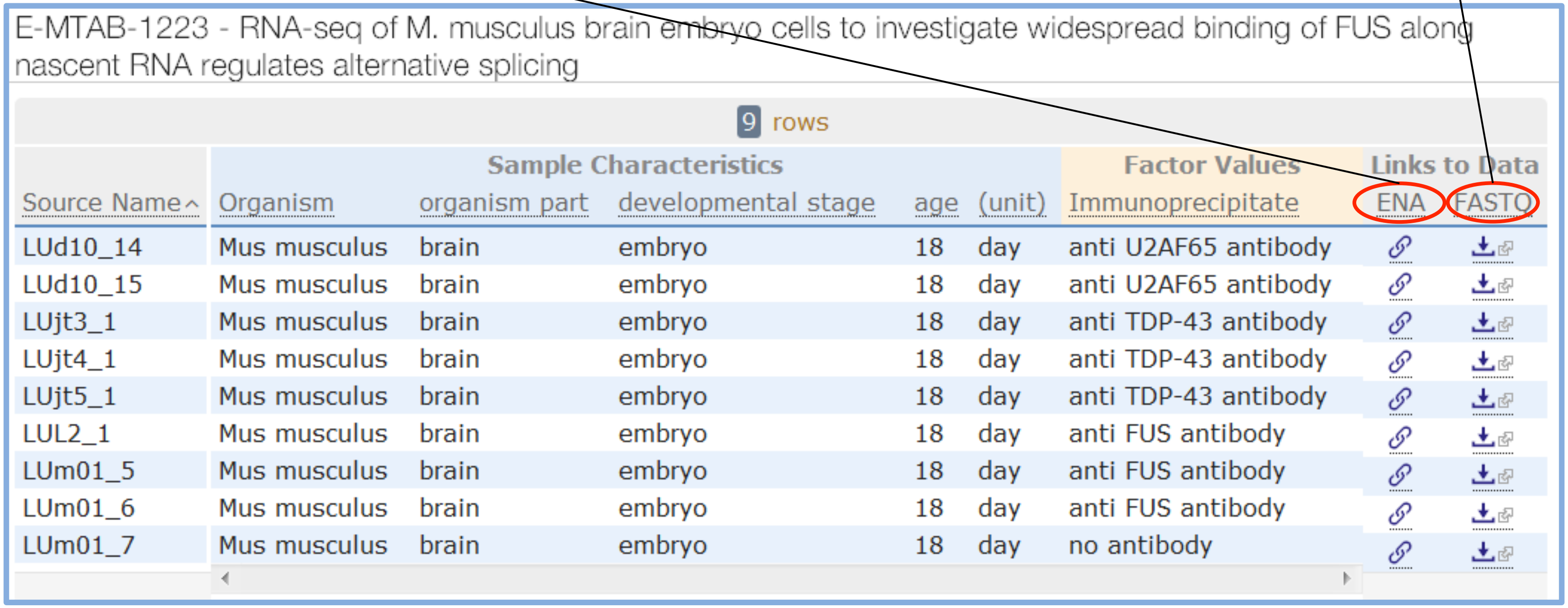
		Sample Characteristics					Factor Values						Links to Data	
Source Name		Organism	OrganismPart	PreservationType	OrganismPa	PreservationTy	dose	(unit)	vehicle	exposure	(unit)	comp	Raw	Processed Matrix
681061186	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	0.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061276	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	0.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061341	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	0.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061410	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	20.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061476	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	20.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061540	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	20.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061602	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	400.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061670	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	400.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061736	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	400.000	ppm	deionised water	52	weeks	potas	⤓	⤓
681061804	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	0.000	ppm	deionised water	100	weeks	potas	⤓	⤓
681061869	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	0.000	ppm	deionised water	100	weeks	potas	⤓	⤓
681061934	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	400.000	ppm	deionised water	100	weeks	potas	⤓	⤓
681062001	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	400.000	ppm	deionised water	100	weeks	potas	⤓	⤓
681062067	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	400.000	ppm	deionised water	100	weeks	potas	⤓	⤓
681062132	s	Rattus norvegicus	kidney	liquid nitrogen/-80C freezer	kidney	liquid nitrogen/-80C freezer	400.000	ppm	deionised water	100	weeks	potas	⤓	⤓

Scroll left and right to see all sample characteristics and factor values

Samples view – RNA-seq experiment

Direct link to European Nucleotide Archive (ENA) record about this sequencing assay

Direct link to fastq files at European Nucleotide Archive (ENA)



E-MTAB-1223 - RNA-seq of M. musculus brain embryo cells to investigate widespread binding of FUS along nascent RNA regulates alternative splicing

9 rows

	Sample Characteristics					Factor Values	Links to Data	
Source Name	Organism	organism part	developmental stage	age	(unit)	Immunoprecipitate	ENA	FASTQ
LUd10_14	Mus musculus	brain	embryo	18	day	anti U2AF65 antibody	🔗	⬇
LUd10_15	Mus musculus	brain	embryo	18	day	anti U2AF65 antibody	🔗	⬇
LUjt3_1	Mus musculus	brain	embryo	18	day	anti TDP-43 antibody	🔗	⬇
LUjt4_1	Mus musculus	brain	embryo	18	day	anti TDP-43 antibody	🔗	⬇
LUjt5_1	Mus musculus	brain	embryo	18	day	anti TDP-43 antibody	🔗	⬇
LUL2_1	Mus musculus	brain	embryo	18	day	anti FUS antibody	🔗	⬇
LUm01_5	Mus musculus	brain	embryo	18	day	anti FUS antibody	🔗	⬇
LUm01_6	Mus musculus	brain	embryo	18	day	anti FUS antibody	🔗	⬇
LUm01_7	Mus musculus	brain	embryo	18	day	no antibody	🔗	⬇

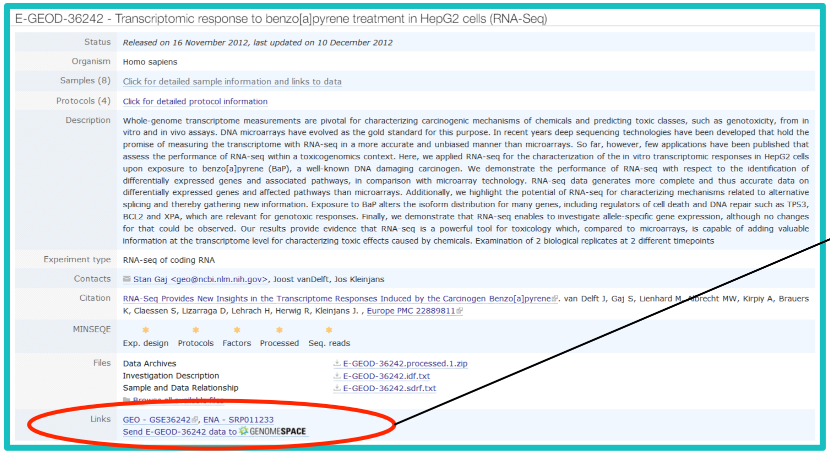
Files and links available for each experiment

Files	Data Archives	E-TOXM-21.processed.1.zip, E-TOXM-21.raw.1.zip
	Investigation Description	E-TOXM-21.idf.txt
	Sample and Data Relationship	E-TOXM-21.sdrf.txt
	Array Design	A-AFFY-25.adf.txt
	R ExpressionSet	E-TOXM-21.eSet.r
	Browse all available files	

ArrayExpress > Experiments > E-TOXM-21

E-TOXM-21 - Transcription profiling of F344/N rats administered with potassium bromate in the drinking water

Status	*Released on 14 November 2006, last updated on 12 October 2011*
Organism	Rattus norvegicus
Samples (15)	Click for detailed sample information and links to data
Array (1)	A-AFFY-25 - Affymetrix GeneChip Rat Expression Array RAE230A [RAE230A]
Protocols (6)	Click for detailed protocol information
Description	Ozone has been proposed for water disinfection because it is more efficient than chlorine for killing microbes and results in much lower levels of carcinogenic trihalomethanes than does chlorination. Ozone leads to formation of hypobromous acid in surface waters with high bromine content and forms brominated organic by-products and bromate. The carcinogenicity and chronic toxicity of potassium bromate (KBrO3) [CAS:7758-02-3;CHEBI:32030] was studied in male B6C3F1 mice and F344/N rats to confirm and extend the results of previous work. Mice were treated with 0, 0.08, 0.4, or 0.8 g/L KBrO3 in the drinking water for up to 100 wk, and rats were provided with 0, 0.02, 0.1, 0.2, or 0.4 g/L KBrO3. Animals were euthanatized, necropsied, and subjected to a complete macroscopic examination. Selected tissues and gross lesions were processed by routine methods for light microscopic examination. The present study showed that KBrO3 is carcinogenic in the rat kidney, thyroid, and mesothelium and is a renal carcinogen in the male mouse, KBrO3 was carcinogenic in rodents at water concentrations as low as 0.02 g/L (20 ppm; 1.5 mg/kg/day). These data can be used to estimate the human health risk that would be associated with changing from chlorination to ozonation for disinfection of drinking water.
Experiment types	transcription profiling by array, compound treatment, dose response
Contact	Geter A David <Geter.David@epamail.epa.gov>
Citation	Kidney Toxicogenomics of Chronic Potassium Bromate Exposure in F344 Male Rats. David R. Geter, William O. Ward, Geremy W. Knapp, Anthony B. DeAngelo, Jessica A. Rubis, Russell D. Owen, James W. Allen, and Don A. Delker.
MIAME	✱ Platforms ✱ Protocols ✱ Factors ✱ Processed ✱ Raw
Files	Data Archives: E-TOXM-21.processed.1.zip, E-TOXM-21.raw.1.zip Investigation Description: E-TOXM-21.idf.txt Sample and Data Relationship: E-TOXM-21.sdrf.txt Array Design: A-AFFY-25.adf.txt R ExpressionSet: E-TOXM-21.eSet.r Browse all available files
Links	Send E-TOXM-21 data to GENOMESPACE



E-GEOD-36242 - Transcriptomic response to benzo[a]pyrene treatment in HepG2 cells (RNA-Seq)

Status	*Released on 16 November 2012, last updated on 10 December 2012*
Organism	Homo sapiens
Samples (8)	Click for detailed sample information and links to data
Protocols (4)	Click for detailed protocol information
Description	Whole-genome transcriptome measurements are pivotal for characterizing carcinogenic mechanisms of chemicals and predicting toxic classes, such as genotoxicity, from in vitro and in vivo assays. DNA microarrays have evolved as the gold standard for this purpose. In recent years deep sequencing technologies have been developed that hold the promise of measuring the transcriptome with RNA-seq in a more accurate and unbiased manner than microarrays. So far, however, few applications have been published that assess the performance of RNA-seq within a toxicogenomics context. Here, we applied RNA-seq for the characterization of the in vitro transcriptomic responses in HepG2 cells upon exposure to benzo[a]pyrene (BaP), a well-known DNA damaging carcinogen. We demonstrate the performance of RNA-seq with respect to the identification of differentially expressed genes and associated pathways, in comparison with microarray technology. RNA-seq data generates more complete and thus accurate data on differentially expressed genes and affected pathways than microarrays. Additionally, we highlight the potential of RNA-seq for characterizing mechanisms related to alternative splicing and thereby gathering new information. Exposure to BaP alters the isoform distribution for many genes, including regulators of cell death and DNA repair such as TP53, BCL2 and XPA, which are relevant for genotoxic responses. Finally, we demonstrate that RNA-seq enables to investigate allele-specific gene expression, although no changes for that could be observed. Our results provide evidence that RNA-seq is a powerful tool for toxicology which, compared to microarrays, is capable of adding valuable information at the transcriptome level for characterizing toxic effects caused by chemicals. Examination of 2 biological replicates at 2 different timepoints
Experiment type	RNA-seq of coding RNA
Contacts	Stan Gaj <geo@ncbi.nlm.nih.gov>, Joost vanDelft, Jos Kleinjans
Citation	RNA-Seq Provides New Insights in the Transcriptome Responses Induced by the Carcinogen Benzo[a]pyrene. van Delft J, Gaj S, Lienhard M, Albrecht MW, Kirpiy A, Brauers K, Claessen S, Lizarraga D, Lehrach H, Herwig R, Kleinjans J. , Europe PMC 22889811
MINSEQE	✱ Exp. design ✱ Protocols ✱ Factors ✱ Processed ✱ Seq. reads
Files	Data Archives: E-GEOD-36242.processed.1.zip Investigation Description: E-GEOD-36242.idf.txt Sample and Data Relationship: E-GEOD-36242.sdrf.txt Browse all available files
Links	GEO - GSE36242, ENA - SRP011233 Send E-GEOD-36242 data to GENOMESPACE

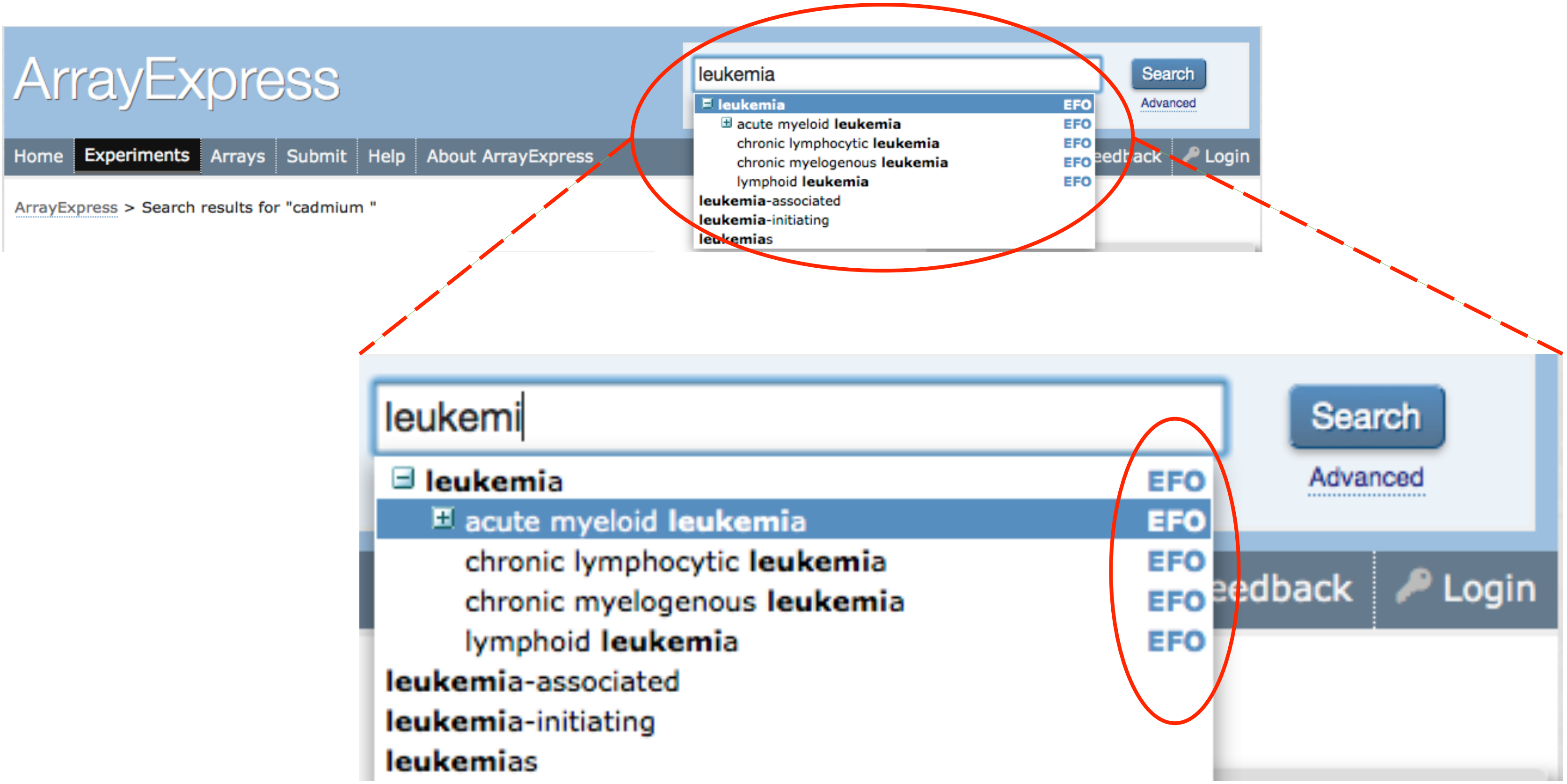
Links	GEO - GSE36242, ENA - SRP011233 Send E-GEOD-36242 data to GENOMESPACE

Links to:

- original NCBI GEO record (GSExxxxx)
- Sequence Read Archive study (SRPxxxxx)

Searching for experiments in ArrayExpress

www.ebi.ac.uk/arrayexpress/experiments/browse.html



Experimental factor ontology (EFO)

http://www.ebi.ac.uk/efo

- Ontology: a way to systematically organise experimental factor terms.

  controlled vocabulary + hierarchy (relationship)

- Used in EBI databases: **BioSD** - BioSample Database

  and external projects (e.g. NHGRI GWAS Catalogue)

- Combine terms from a subset of well-maintained and compatible ontologies, e.g.

  Gene Ontology (cellular component + biological process terms)
  NCBI Taxonomy

Building EFO - an example

Take all experimental factors

- sarcoma
- cancer
- neoplasm
- disease
- Kaposi's sarcoma

Find the logical connection between them

disease is the parent term

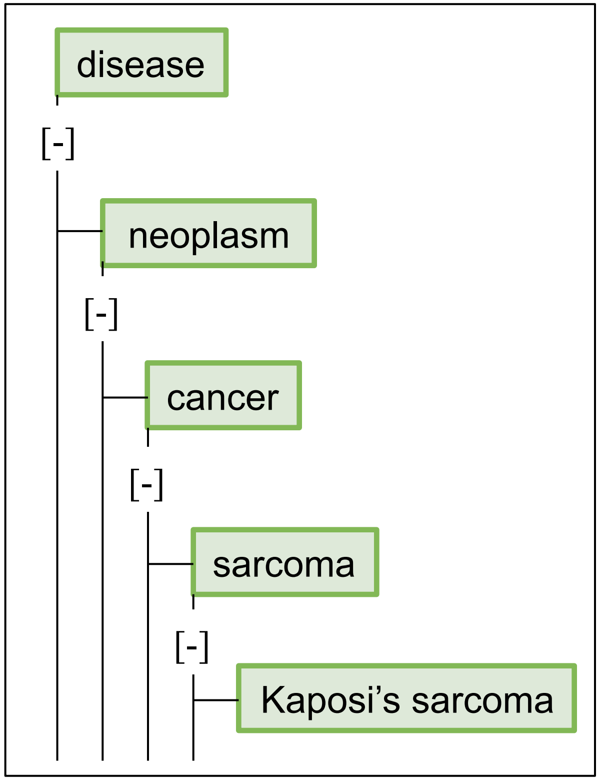
neoplasm is a type of disease

cancer is synonym of neoplasm

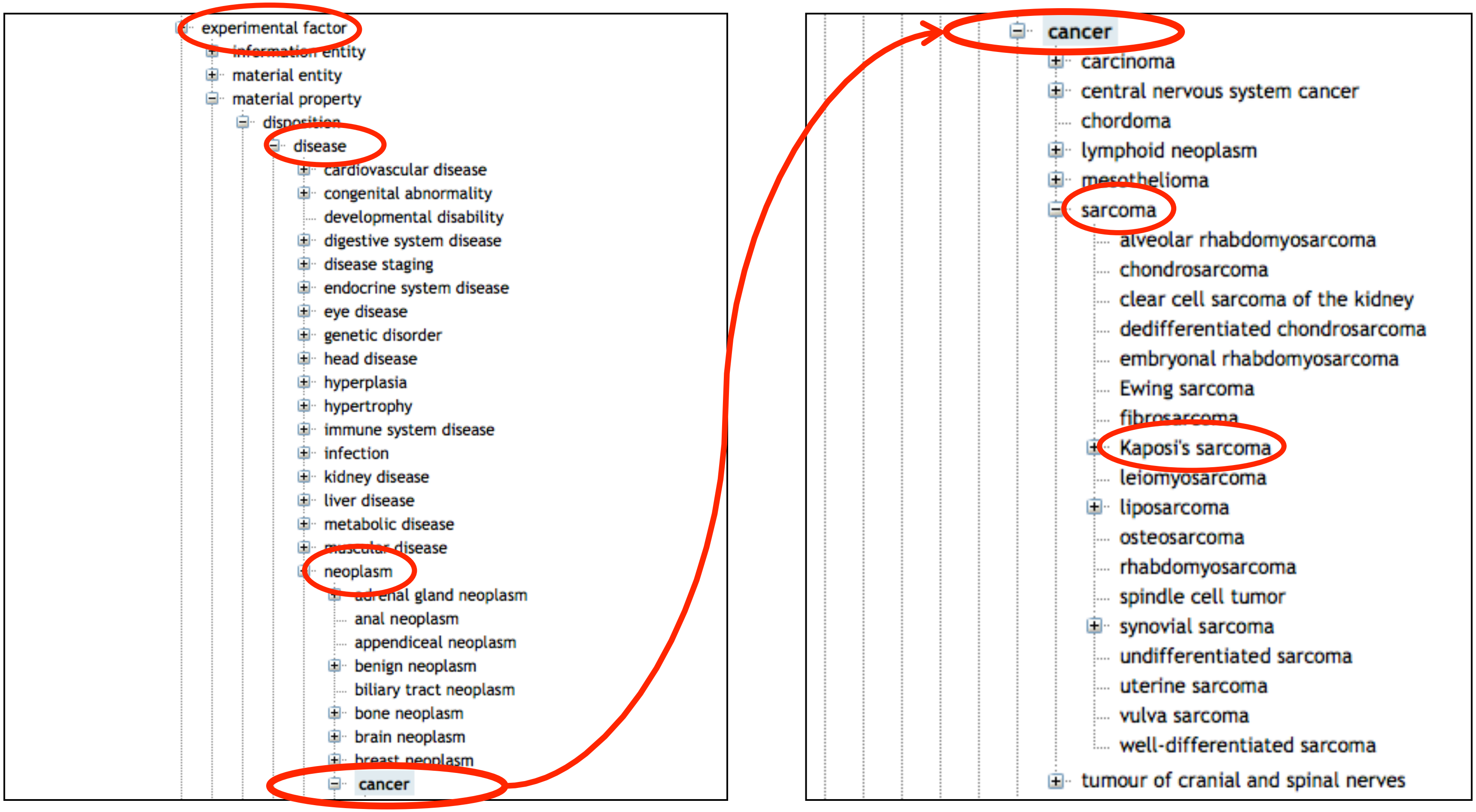
sarcoma is a type of cancer

Kaposi's sarcoma is a type of sarcoma

Organize them in an ontology



Exploring EFO - an example



- experimental factor
  - information entity
  - material entity
  - material property
    - disposition
      - disease
        - cardiovascular disease
        - congenital abnormality
        - developmental disability
        - digestive system disease
        - disease staging
        - endocrine system disease
        - eye disease
        - genetic disorder
        - head disease
        - hyperplasia
        - hypertrophy
        - immune system disease
        - infection
        - kidney disease
        - liver disease
        - metabolic disease
        - muscular disease
        - neoplasm
          - adrenal gland neoplasm
          - anal neoplasm
          - appendiceal neoplasm
          - benign neoplasm
          - biliary tract neoplasm
          - bone neoplasm
          - brain neoplasm
          - breast neoplasm
          - cancer

- cancer
  - carcinoma
  - central nervous system cancer
  - chordoma
  - lymphoid neoplasm
  - mesothelioma
  - sarcoma
    - alveolar rhabdomyosarcoma
    - chondrosarcoma
    - clear cell sarcoma of the kidney
    - dedifferentiated chondrosarcoma
    - embryonal rhabdomyosarcoma
    - Ewing sarcoma
    - fibrosarcoma
    - Kaposi's sarcoma
    - leiomyosarcoma
    - liposarcoma
    - osteosarcoma
    - rhabdomyosarcoma
    - spindle cell tumor
    - synovial sarcoma
    - undifferentiated sarcoma
    - uterine sarcoma
    - vulva sarcoma
    - well-differentiated sarcoma
  - tumour of cranial and spinal nerves

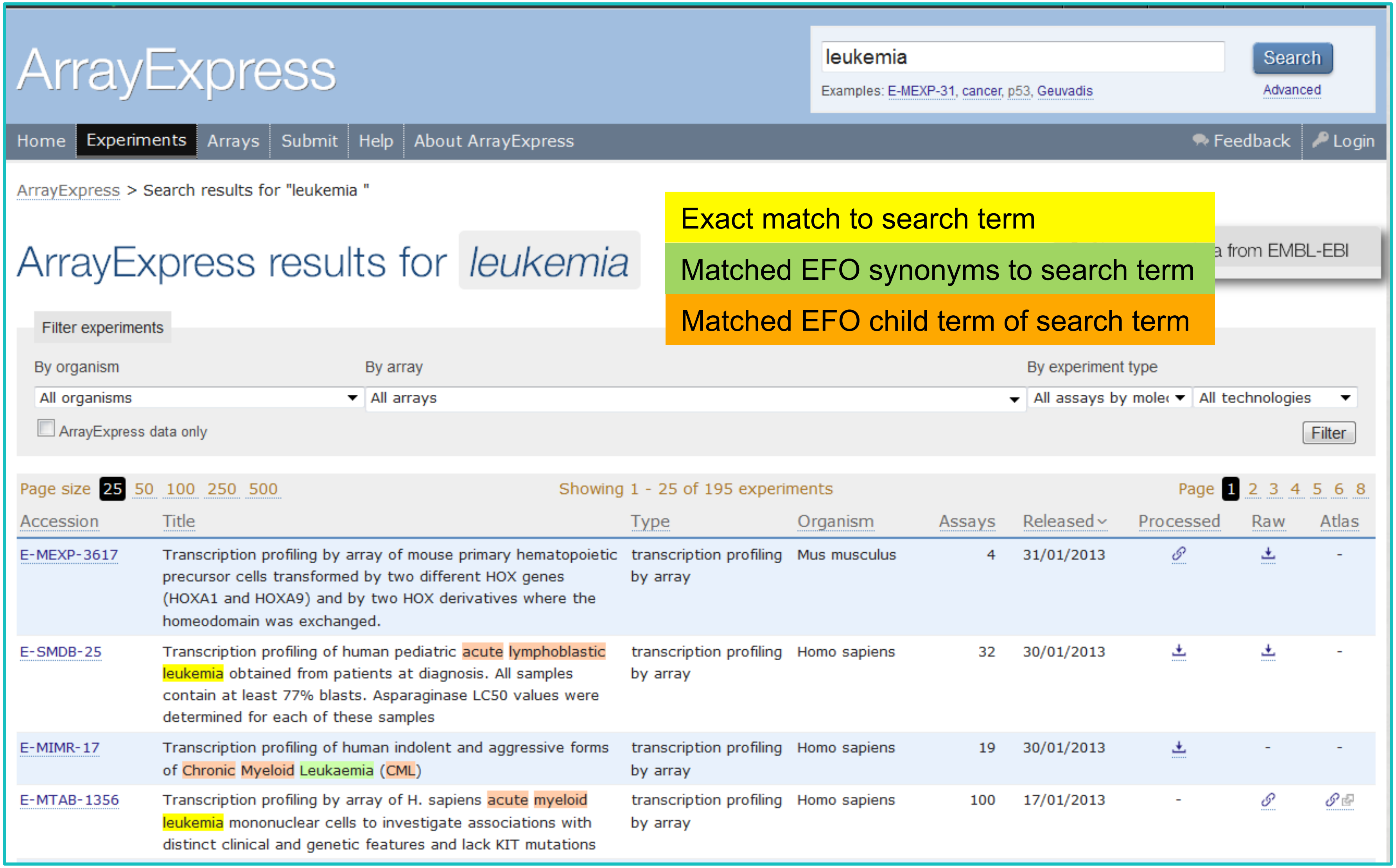
Experimental factor ontology (EFO)

http://www.ebi.ac.uk/efo

EFO developed to:

- increase the richness of annotations in databases
- expand on search terms when querying ArrayExpress and Expression Atlas
  - using synonyms (e.g. “cerebral cortex” = “adult brain cortex”)
  - using child terms (e.g. “bone” → “rib” and “vertebra”)
- promote consistency (e.g. F/female/, 1day/24hours)
- facilitate automatic annotation and integration of external data (e.g. changing “gender” to “sex” automatically)

Example search: “leukemia”



ArrayExpress > Search results for "leukemia "

ArrayExpress results for *leukemia*

Exact match to search term

Matched EFO synonyms to search term

Matched EFO child term of search term

a from EMBL-EBI

Filter experiments

By organism: All organisms

By array: All arrays

By experiment type: All assays by mole(| All technologies

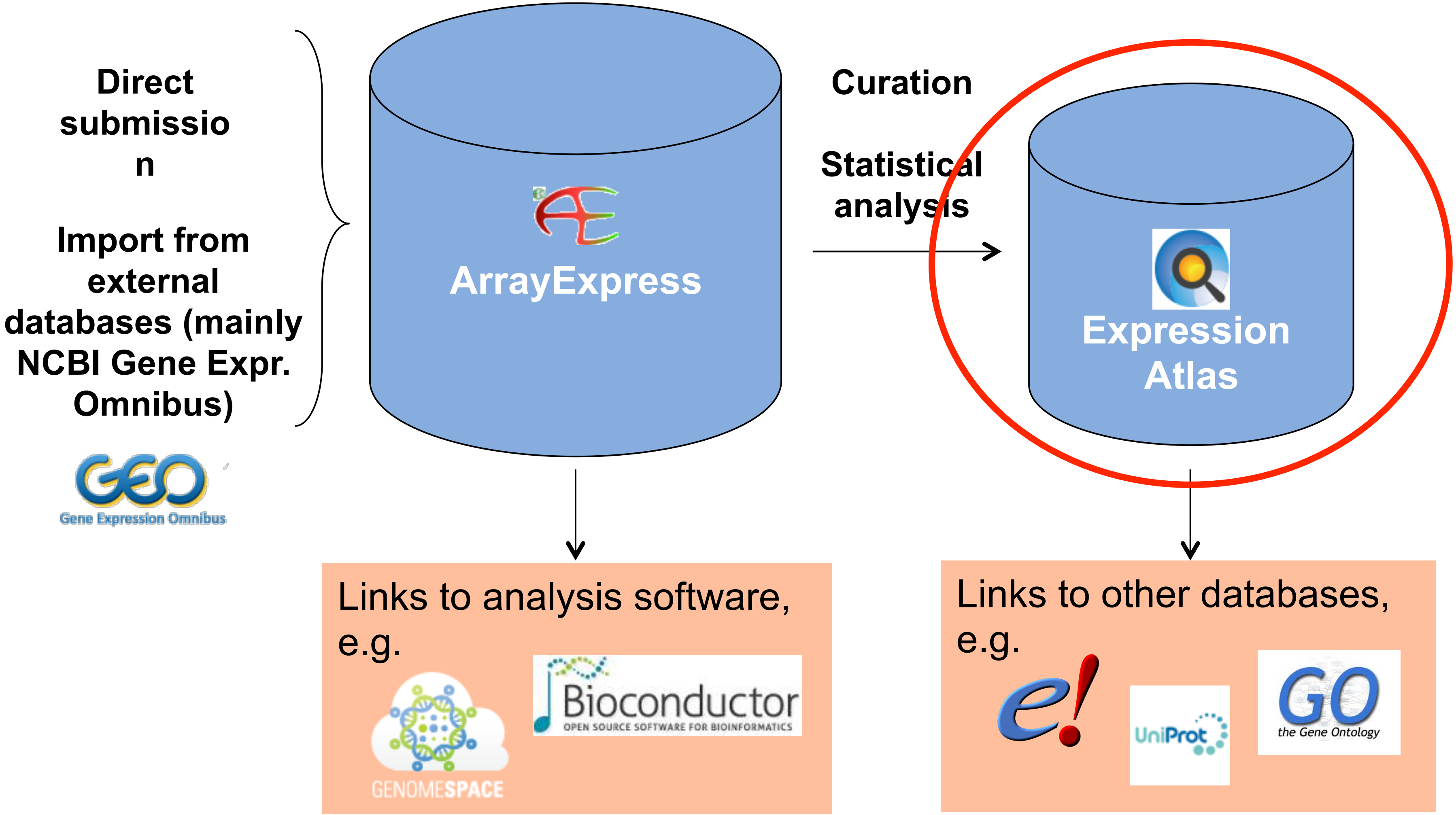
☐ ArrayExpress data only

Filter

Page size **25** 50 100 250 500 | Showing 1 - 25 of 195 experiments | Page **1** 2 3 4 5 6 8

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-MEXP-3617	Transcription profiling by array of mouse primary hematopoietic precursor cells transformed by two different HOX genes (HOXA1 and HOXA9) and by two HOX derivatives where the homeodomain was exchanged.	transcription profiling by array	Mus musculus	4	31/01/2013	🔗	⬇	-
E-SMDB-25	Transcription profiling of human pediatric acute lymphoblastic leukemia obtained from patients at diagnosis. All samples contain at least 77% blasts. Asparaginase LC50 values were determined for each of these samples	transcription profiling by array	Homo sapiens	32	30/01/2013	⬇	⬇	-
E-MIMR-17	Transcription profiling of human indolent and aggressive forms of Chronic Myeloid Leukaemia (CML)	transcription profiling by array	Homo sapiens	19	30/01/2013	⬇	-	-
E-MTAB-1356	Transcription profiling by array of H. sapiens acute myeloid leukemia mononuclear cells to investigate associations with distinct clinical and genetic features and lack KIT mutations	transcription profiling by array	Homo sapiens	100	17/01/2013	-	🔗	🔗 ⧉

The two databases: how are they related?



Direct submission

Import from external databases (mainly NCBI Gene Expr. Omnibus)

ArrayExpress

Curation

Statistical analysis

Expression Atlas

Links to analysis software, e.g.

Links to other databases, e.g.

Expression Atlas – when to use it?

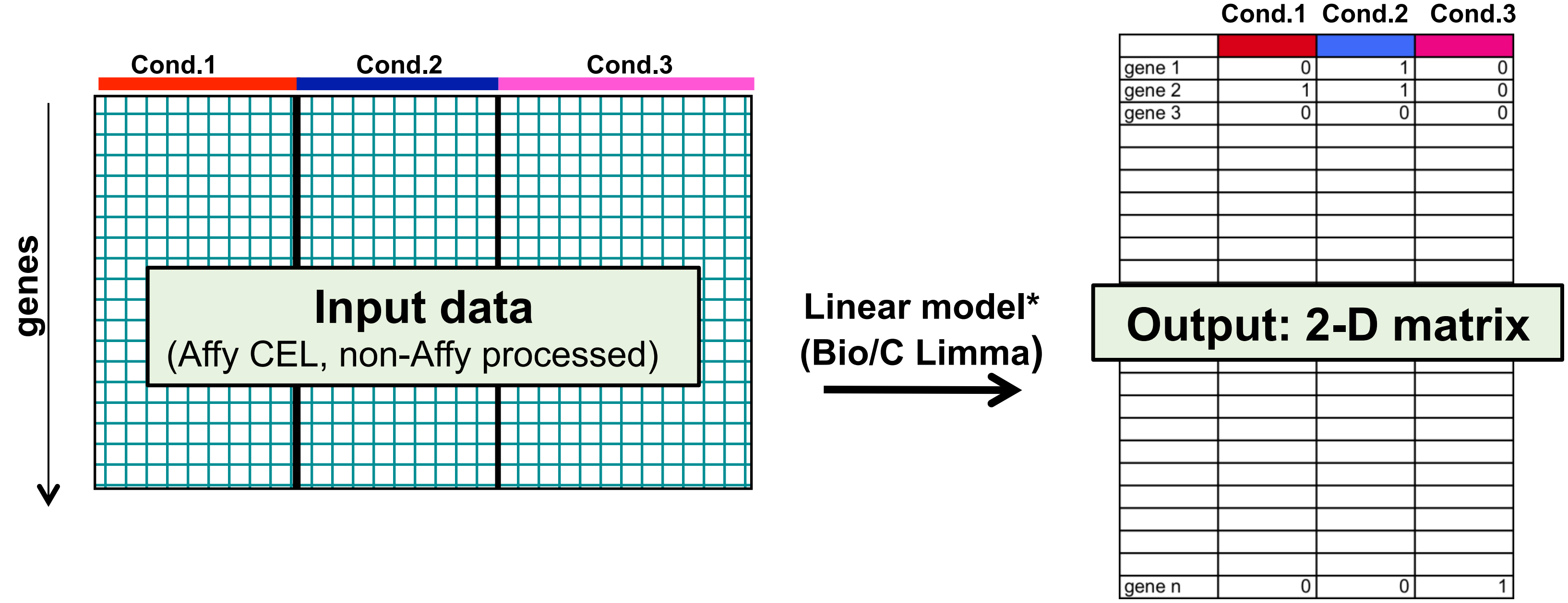
- Find out if the expression of a gene (or a group of genes with a common gene attribute, e.g. GO term) change(s) across all the experiments available in the Expression Atlas;

- Discover which genes are differentially expressed in a particular biological condition that you are interested in.

Expression Atlas construction

Analysis pipeline

A dummy example:



Cond.1 Cond.2 Cond.3

genes

Input data
(Affy CEL, non-Affy processed)

Linear model*
(Bio/C Limma)

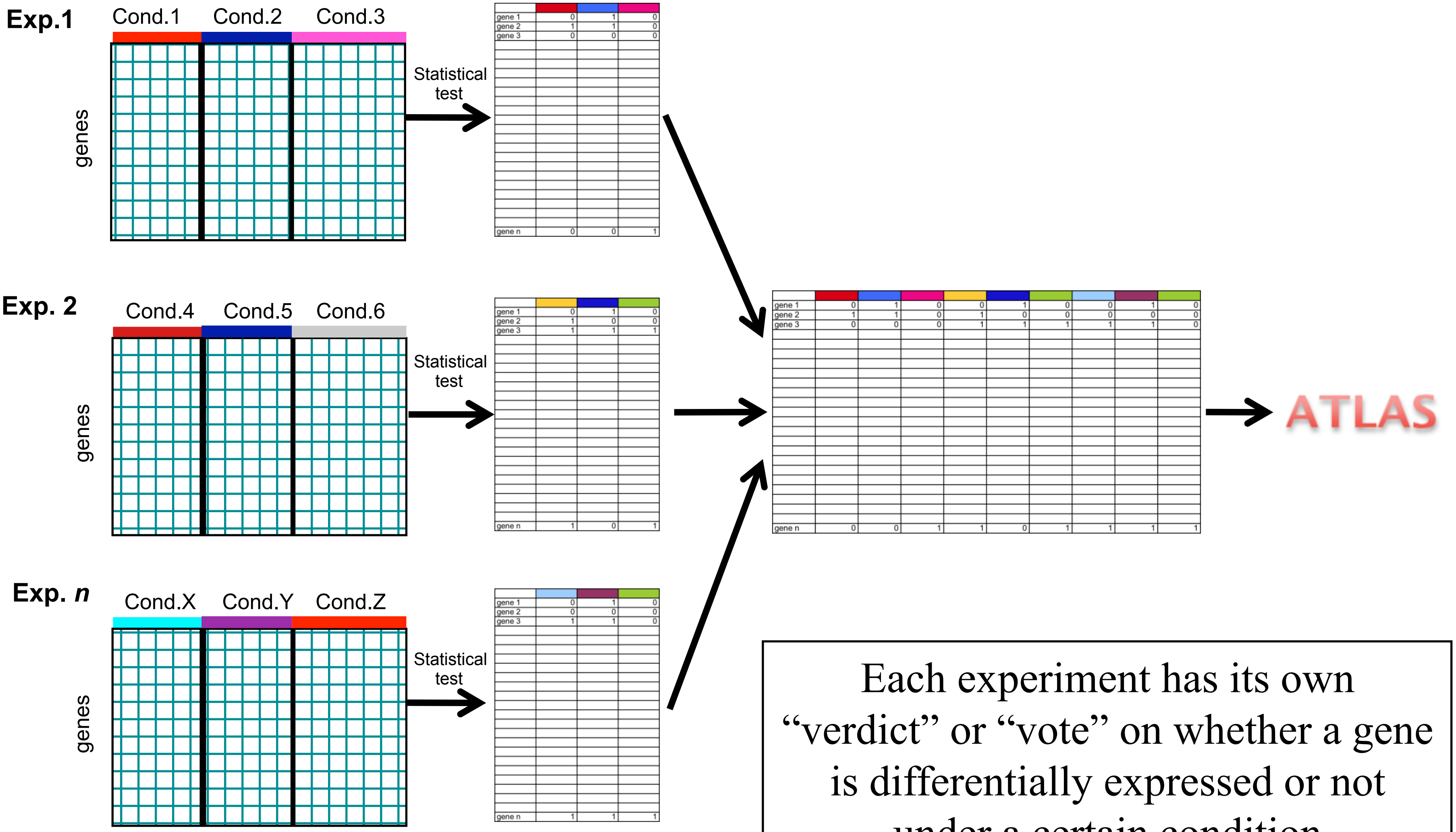
Output: 2-D matrix

	Cond.1	Cond.2	Cond.3
gene 1	0	1	0
gene 2	1	1	0
gene 3	0	0	0
gene n	0	0	1

1= differentially expressed
0 = not differentially expressed

* More information about the statistical methodology:
http://nar.oxfordjournals.org/content/38/suppl_1/D690.full

Expression Atlas construction



	🟥	🟦	🟪
gene 1	0	1	0
gene 2	1	1	0
gene 3	0	0	0
gene n	0	0	1

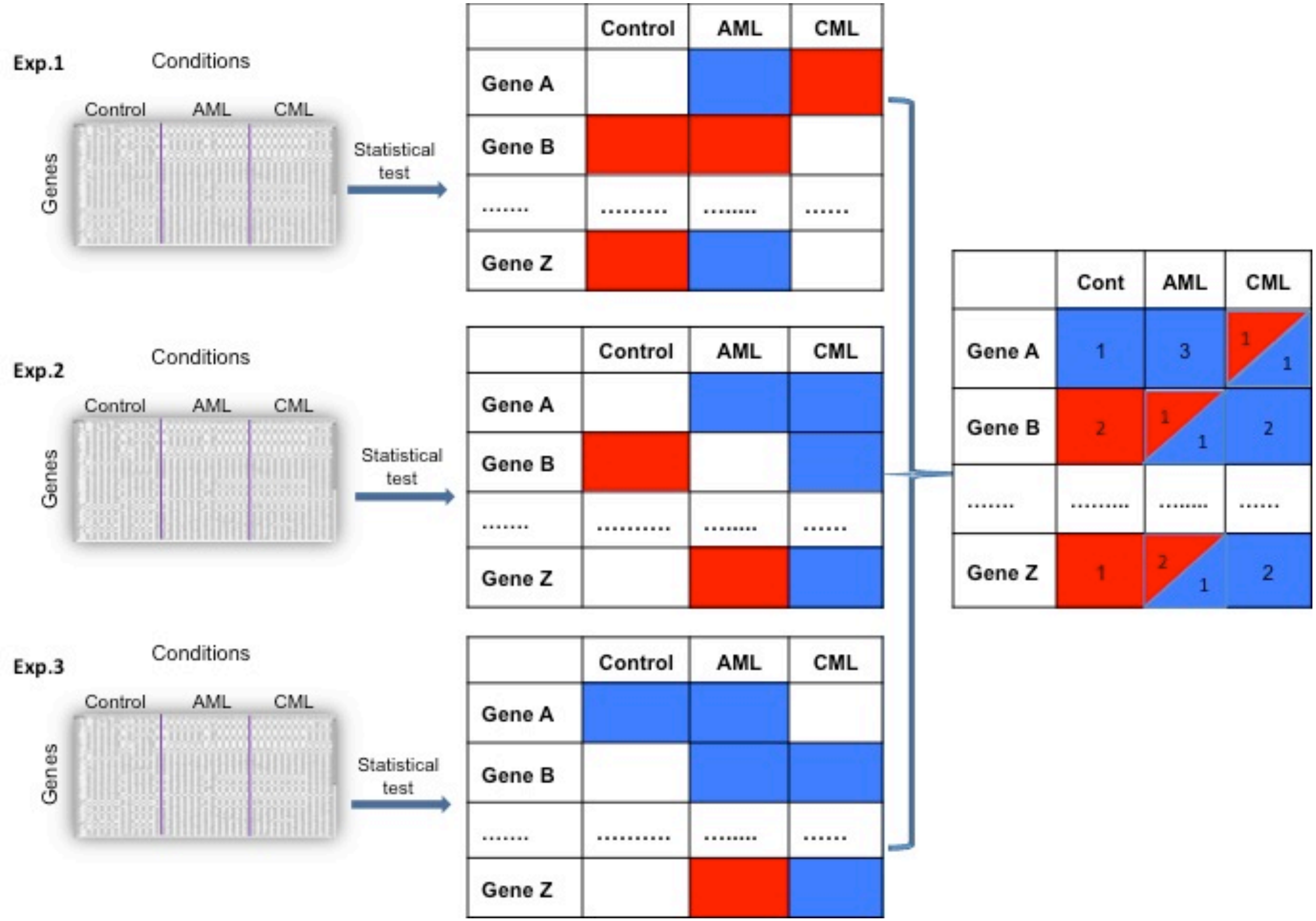
	🟨	🟦	🟩
gene 1	0	1	0
gene 2	1	0	0
gene 3	1	1	1
gene n	1	0	1

	🟦	🟪	🟩
gene 1	0	1	0
gene 2	0	0	0
gene 3	1	1	0
gene n	1	1	1

	🟥	🟦	🟪	🟨	🟦	🟩	🟦	🟪	🟩
gene 1	0	1	0	0	1	0	0	1	0
gene 2	1	1	0	1	0	0	0	0	0
gene 3	0	0	0	1	1	1	1	1	0
gene n	0	0	1	1	0	1	1	1	1

Each experiment has its own “verdict” or “vote” on whether a gene is differentially expressed or not under a certain condition

Expression Atlas construction



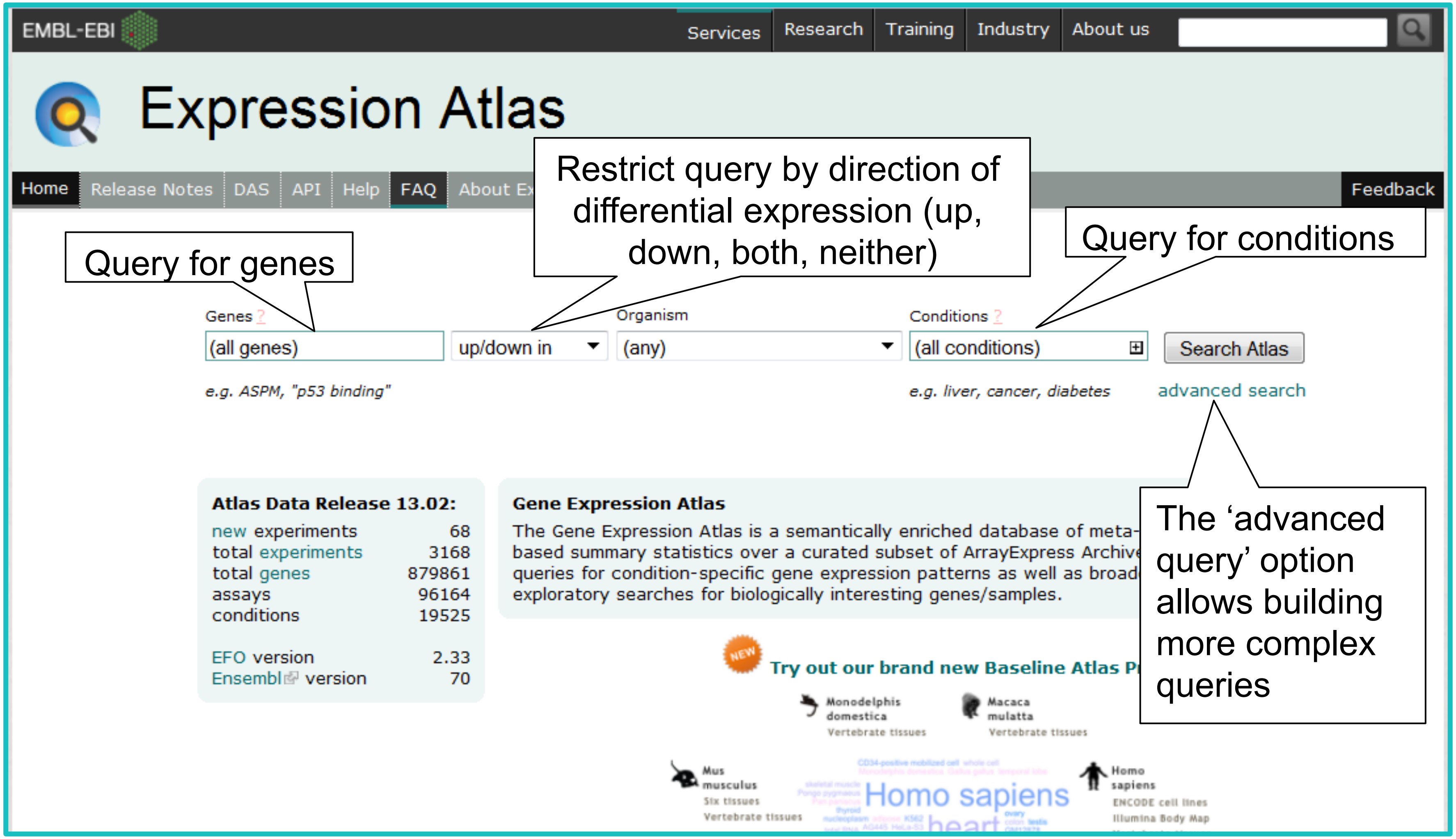
	Control	AML	CML
Gene A		🟦	🟥
Gene B	🟥	🟥	
....	....	....	......
Gene Z	🟥	🟦	

	Control	AML	CML
Gene A		🟦	🟦
Gene B	🟥		🟦
....	....	....	......
Gene Z		🟥	🟦

	Control	AML	CML
Gene A	🟦	🟦	
Gene B		🟦	🟦
....	....	....	......
Gene Z		🟥	🟦

	Cont	AML	CML
Gene A	🟦 1	🟦 3	🟥 1 🟦 1
Gene B	🟥 2	🟥 1 🟦 1	🟦 2
....	....	....	......
Gene Z	🟥 1	🟥 2 🟦 1	🟦 2

Expression Atlas home page http://www.ebi.ac.uk/gxa



Atlas Data Release 13.02:

new experiments	68
total experiments	3168
total genes	879861
assays	96164
conditions	19525
EFO version	2.33
Ensembl version	70

Gene Expression Atlas

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

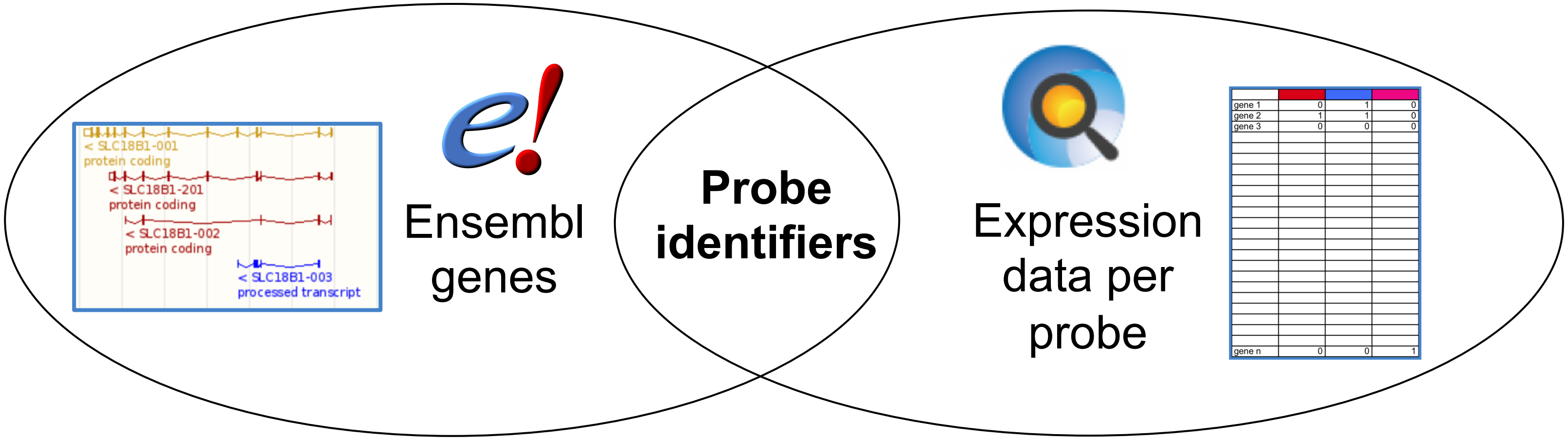
Query for genes

Restrict query by direction of differential expression (up, down, both, neither)

Query for conditions

The 'advanced query' option allows building more complex queries

Mapping microarray probes to genes



Ensembl genes

Probe identifiers

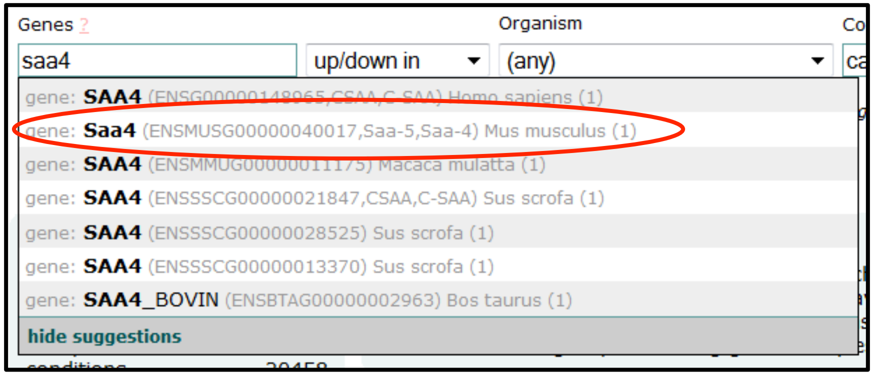
Expression data per probe

- Every (~monthly) Atlas release takes the latest Ensembl gene – probe identifier mapping data.
- From Ensembl genes, we also get:
  - Compara genes
  - External references (xrefs) to other databases

    E.g. UniProt protein IDs, NCBI RefSeq IDs, HGNC gene symbols, gene ontology terms, InterPro terms

Expression Atlas home page

The ‘Genes’ and ‘Conditions’ search boxes



saa4

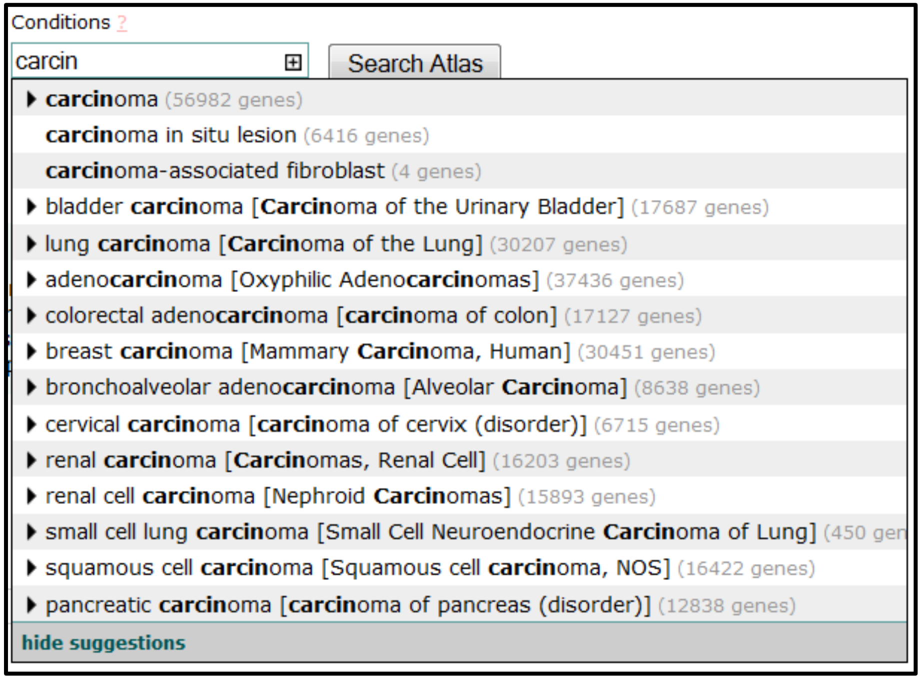
up/down in

(any)

gene: **SAA4** (ENSG00000148965,CSAA,C-SAA) Homo sapiens (1)
gene: **Saa4** (ENSMUSG00000040017,Saa-5,Saa-4) Mus musculus (1)
gene: **SAA4** (ENSMMUG00000011175) Macaca mulatta (1)
gene: **SAA4** (ENSSSCG00000021847,CSAA,C-SAA) Sus scrofa (1)
gene: **SAA4** (ENSSSCG00000028525) Sus scrofa (1)
gene: **SAA4** (ENSSSCG00000013370) Sus scrofa (1)
gene: **SAA4**_BOVIN (ENSBTAG00000002963) Bos taurus (1)

Genes

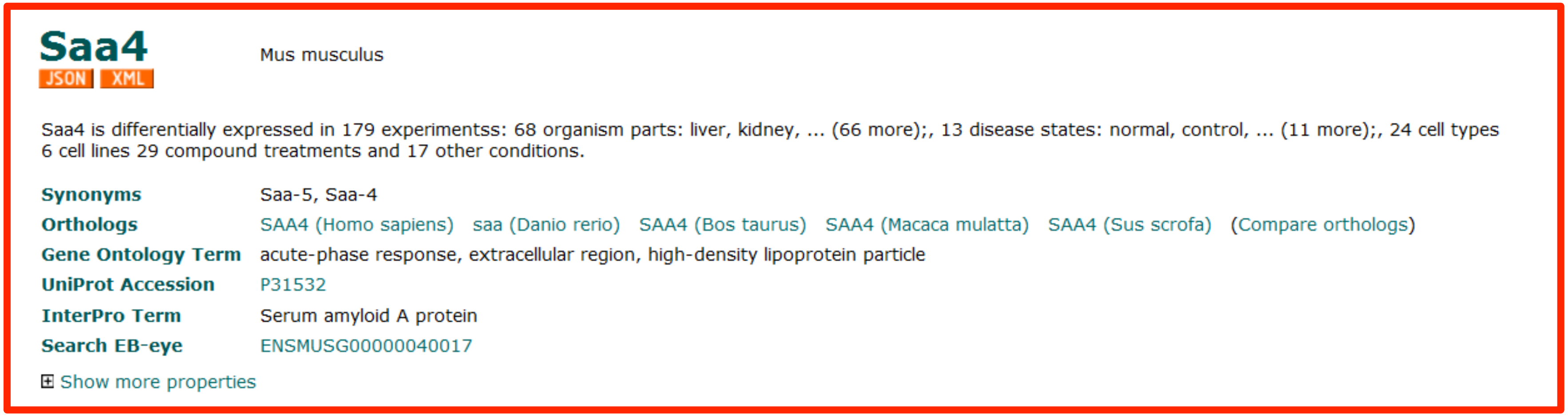
Conditions



carcin

carcinoma (56982 genes)
carcinoma in situ lesion (6416 genes)
carcinoma-associated fibroblast (4 genes)
bladder **carcin**oma [**Carcin**oma of the Urinary Bladder] (17687 genes)
lung **carcin**oma [**Carcin**oma of the Lung] (30207 genes)
adeno**carcin**oma [Oxyphilic Adeno**carcin**omas] (37436 genes)
colorectal adeno**carcin**oma [**carcin**oma of colon] (17127 genes)
breast **carcin**oma [Mammary **Carcin**oma, Human] (30451 genes)
bronchoalveolar adeno**carcin**oma [Alveolar **Carcin**oma] (8638 genes)
cervical **carcin**oma [**carcin**oma of cervix (disorder)] (6715 genes)
renal **carcin**oma [**Carcin**omas, Renal Cell] (16203 genes)
renal cell **carcin**oma [Nephroid **Carcin**omas] (15893 genes)
small cell lung **carcin**oma [Small Cell Neuroendocrine **Carcin**oma of Lung] (450 gen
squamous cell **carcin**oma [Squamous cell **carcin**oma, NOS] (16422 genes)
pancreatic **carcin**oma [**carcin**oma of pancreas (disorder)] (12838 genes)

Expression Atlas - single gene query



Saa4

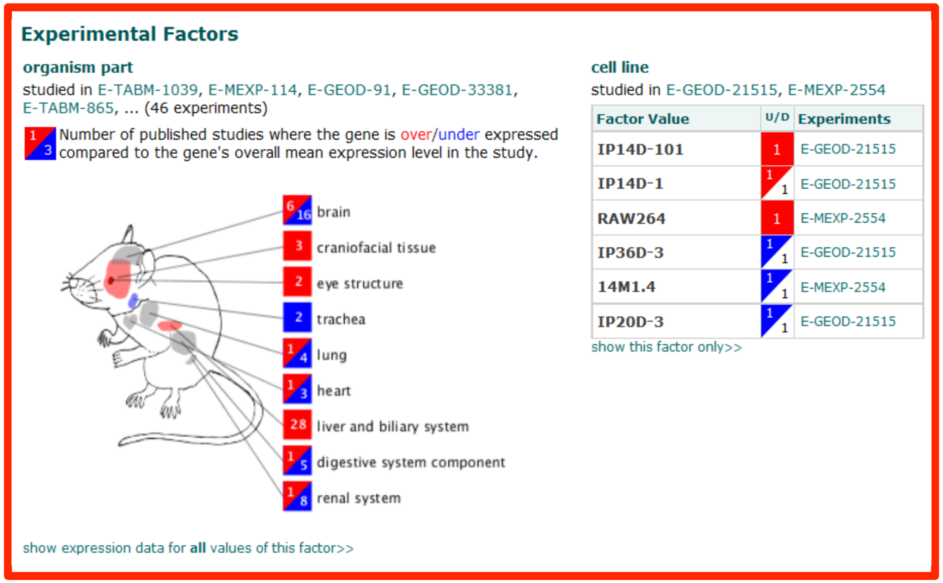
JSON XML

Mus musculus

Saa4 is differentially expressed in 179 experimentss: 68 organism parts: liver, kidney, ... (66 more);, 13 disease states: normal, control, ... (11 more);, 24 cell types 6 cell lines 29 compound treatments and 17 other conditions.

Synonyms	Saa-5, Saa-4
Orthologs	SAA4 (Homo sapiens) saa (Danio rerio) SAA4 (Bos taurus) SAA4 (Macaca mulatta) SAA4 (Sus scrofa) (Compare orthologs)
Gene Ontology Term	acute-phase response, extracellular region, high-density lipoprotein particle
UniProt Accession	P31532
InterPro Term	Serum amyloid A protein
Search EB-eye	ENSMUSG00000040017

⊞ Show more properties



Experimental Factors

organism part

studied in E-TABM-1039, E-MEXP-114, E-GEOD-91, E-GEOD-33381, E-TABM-865, ... (46 experiments)

1 3 Number of published studies where the gene is over/under expressed compared to the gene's overall mean expression level in the study.

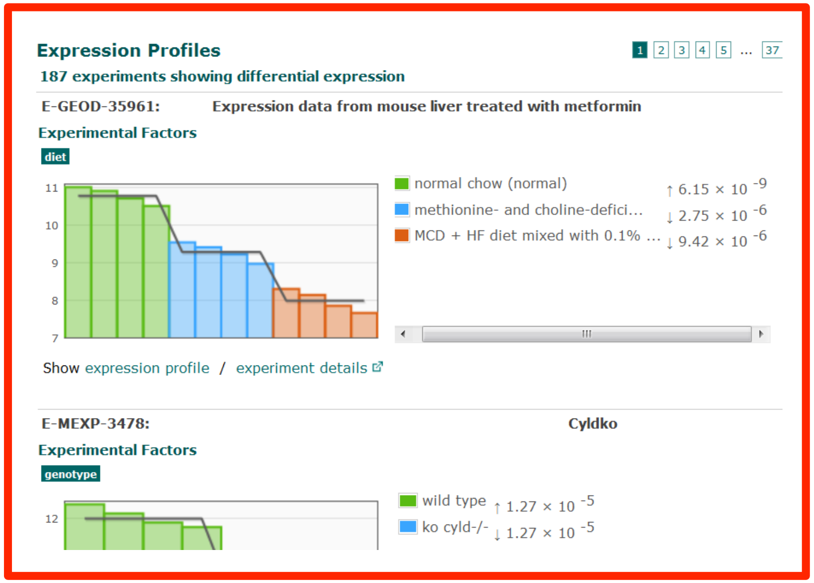
show expression data for **all** values of this factor>>

cell line

studied in E-GEOD-21515, E-MEXP-2554

Factor Value	U/D	Experiments
IP14D-101	1	E-GEOD-21515
IP14D-1	1 1	E-GEOD-21515
RAW264	1	E-MEXP-2554
IP36D-3	1 1	E-GEOD-21515
14M1.4	1 1	E-MEXP-2554
IP20D-3	1 1	E-GEOD-21515

show this factor only>>



Expression Profiles

187 experiments showing differential expression

E-GEOD-35961: **Expression data from mouse liver treated with metformin**

Experimental Factors

diet

- 🟩 normal chow (normal) ↑ 6.15×10^{-9}
- 🟦 methionine- and choline-defici... ↓ 2.75×10^{-6}
- 🟧 MCD + HF diet mixed with 0.1% ... ↓ 9.42×10^{-6}

Show expression profile / experiment details

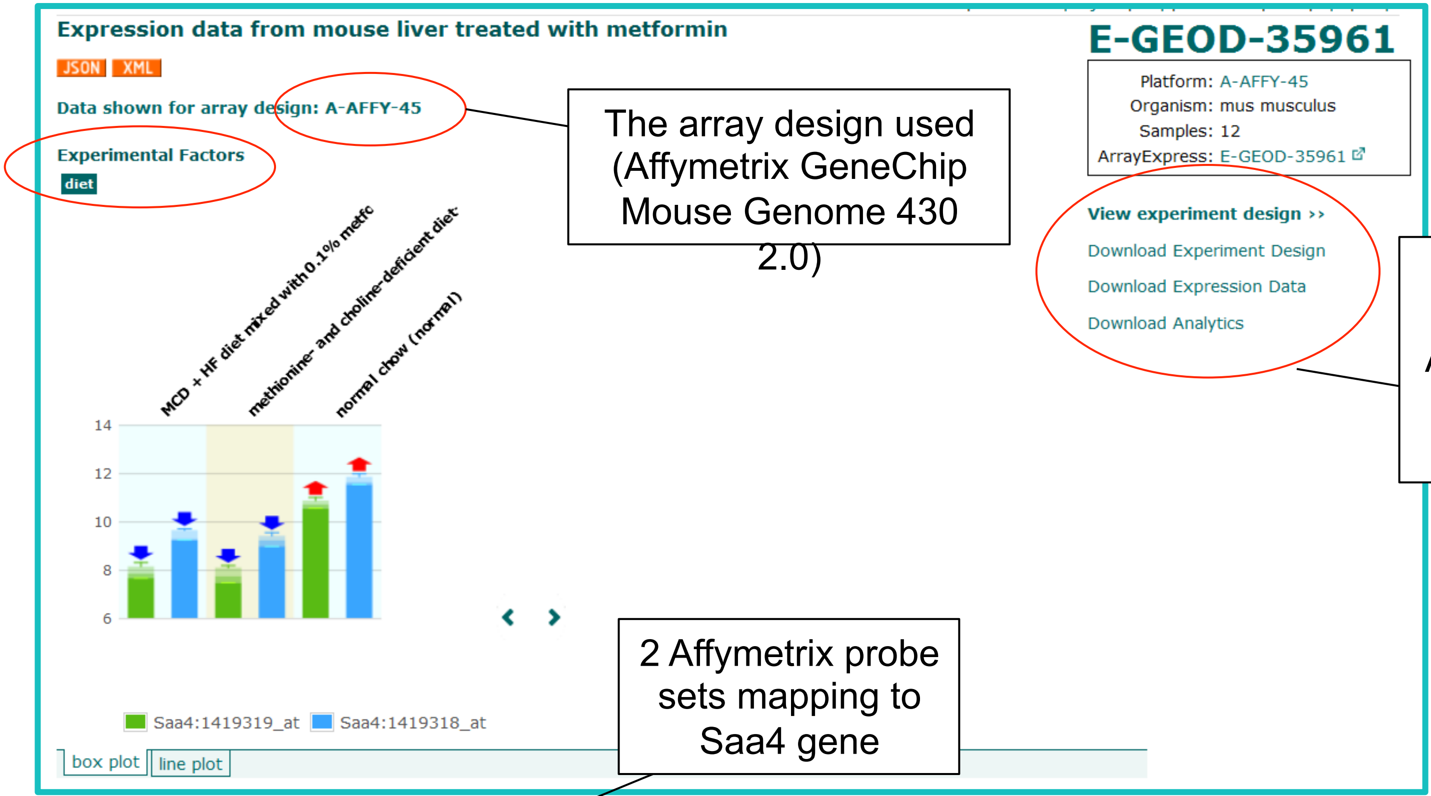
E-MEXP-3478: **Cyldko**

Experimental Factors

genotype

- 🟩 wild type ↑ 1.27×10^{-5}
- 🟦 ko cyld-/- ↓ 1.27×10^{-5}

Expression Atlas - single gene query



Expression data from mouse liver treated with metformin

JSON XML

Data shown for array design: A-AFFY-45

Experimental Factors

diet

E-GEOD-35961

Platform: A-AFFY-45
Organism: mus musculus
Samples: 12
ArrayExpress: E-GEOD-35961

View experiment design ››

Download Experiment Design

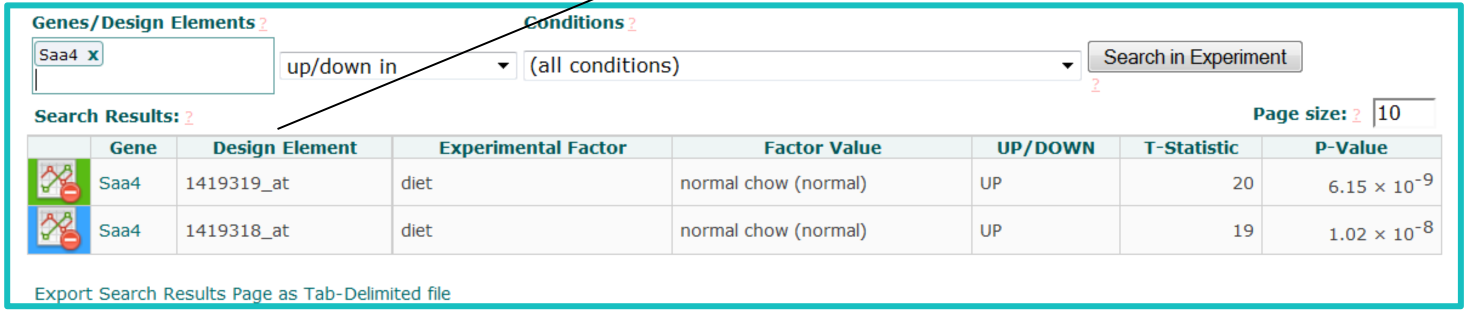
Download Expression Data

Download Analytics

The array design used (Affymetrix GeneChip Mouse Genome 430 2.0)

Experiment design (similar to ArrayExpress samples view)
Analytics = statistics

2 Affymetrix probe sets mapping to Saa4 gene



Genes/Design Elements ?

Conditions ?

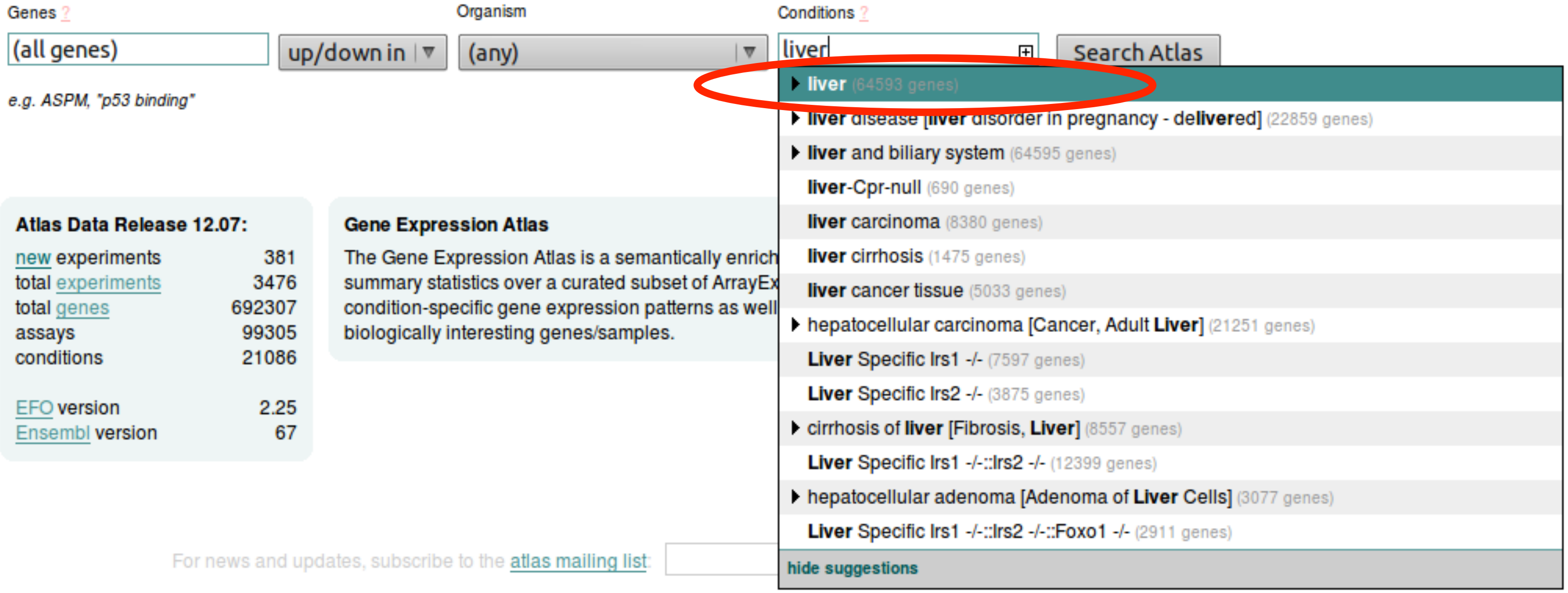
Search Results: ?

Page size: ? 10

	Gene	Design Element	Experimental Factor	Factor Value	UP/DOWN	T-Statistic	P-Value
	Saa4	1419319_at	diet	normal chow (normal)	UP	20	6.15×10^{-9}
	Saa4	1419318_at	diet	normal chow (normal)	UP	19	1.02×10^{-8}

Export Search Results Page as Tab-Delimited file

Atlas ‘condition-only’ query



Genes: (all genes)

up/down in

Organism: (any)

Conditions: liver

e.g. ASPM, "p53 binding"

- liver (64593 genes)
- liver disease [liver disorder in pregnancy - delivered] (22859 genes)
- liver and biliary system (64595 genes)
- liver-Cpr-null (690 genes)
- liver carcinoma (8380 genes)
- liver cirrhosis (1475 genes)
- liver cancer tissue (5033 genes)
- hepatocellular carcinoma [Cancer, Adult Liver] (21251 genes)
- Liver Specific Irs1 -/- (7597 genes)
- Liver Specific Irs2 -/- (3875 genes)
- cirrhosis of liver [Fibrosis, Liver] (8557 genes)
- Liver Specific Irs1 -/-::Irs2 -/- (12399 genes)
- hepatocellular adenoma [Adenoma of Liver Cells] (3077 genes)
- Liver Specific Irs1 -/-::Irs2 -/-::Foxo1 -/- (2911 genes)

Atlas Data Release 12.07:

new experiments	381
total experiments	3476
total genes	692307
assays	99305
conditions	21086
EFO version	2.25
Ensembl version	67

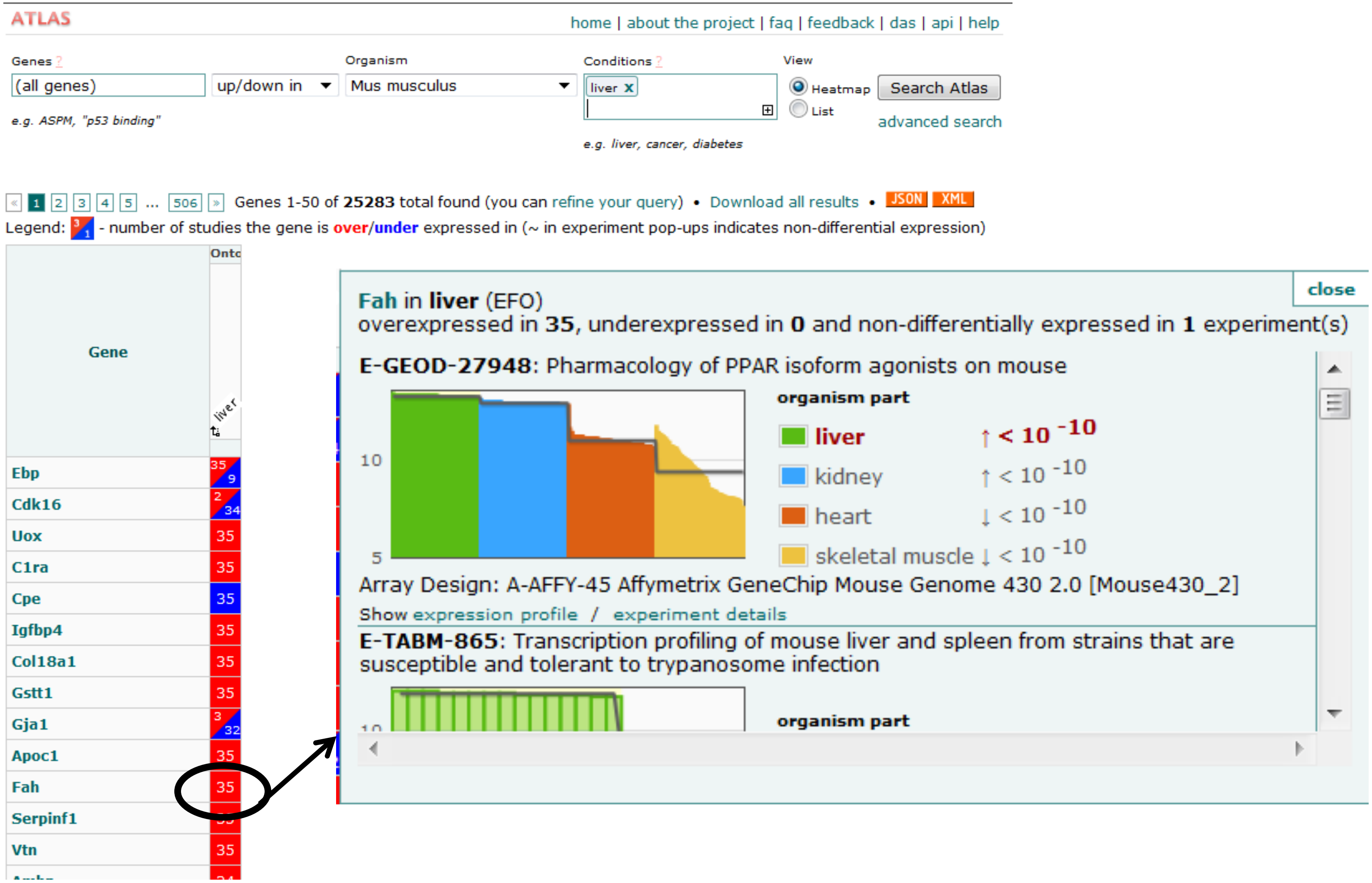
Gene Expression Atlas

The Gene Expression Atlas is a semantically enrich summary statistics over a curated subset of ArrayEx condition-specific gene expression patterns as well biologically interesting genes/samples.

For news and updates, subscribe to the atlas mailing list:

Atlas ‘condition-only’ query (cont’d)

heatmap view



ATLAS

home | about the project | faq | feedback | das | api | help

Genes ?: (all genes)

up/down in

Organism: Mus musculus

Conditions ?: liver x

View: ● Heatmap ○ List

Search Atlas

advanced search

e.g. ASPM, "p53 binding"

e.g. liver, cancer, diabetes

« 1 2 3 4 5 ... 506 » Genes 1-50 of **25283** total found (you can refine your query) • Download all results • JSON XML

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

Gene	liver
Ebp	35 9
Cdk16	2 34
Uox	35
C1ra	35
Cpe	35
Igfbp4	35
Col18a1	35
Gstt1	35
Gja1	3 32
Apoc1	35
Fah	35
Serpinf1	35
Vtn	35

close

Fah in **liver** (EFO)

overexpressed in **35**, underexpressed in **0** and non-differentially expressed in **1** experiment(s)

E-GEOD-27948: Pharmacology of PPAR isoform agonists on mouse

organism part

- 🟩 **liver** ↑ $< 10^{-10}$
- 🟦 kidney ↑ $< 10^{-10}$
- 🟧 heart ↓ $< 10^{-10}$
- 🟨 skeletal muscle ↓ $< 10^{-10}$

Array Design: A-AFFY-45 Affymetrix GeneChip Mouse Genome 430 2.0 [Mouse430_2]

Show expression profile / experiment details

E-TABM-865: Transcription profiling of mouse liver and spleen from strains that are susceptible and tolerant to trypanosome infection

organism part

Atlas gene + condition query

Genes ?
Wnt receptor signali... ✕
up/down in ▾
Organism
Homo sapiens ▾
Conditions ?
carcinoma ✕ ⊞
View
● Heatmap
○ List
Search Atlas
advanced search

e.g. ASPM, "p53 binding"

e.g. liver, cancer, diabetes

« **1** 2 3 4 5 6 » Genes 1-50 of **288** total found (you can refine your query) • Download all results • JSON XML

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

Gene	Ontology																			
	carcinoma	adenocarcinoma	adrenocortical carcinoma	bladder carcinoma	breast carcinoma	cervical carcinoma	endometrioid carcinoma	esophageal carcinoma	gastric carcinoma	hepatocellular carcinoma	hereditary leiomyomatosis an...	large cell neuroendocrine car...	lung carcinoma	melanoma	ovarian carcinoma	pancreatic carcinoma	prostate carcinoma	renal carcinoma	squamous cell carcinoma	thyroid carcinoma
COL1A1	🟥34 🟦19	🟥9 🟦6	🟥2	🟥4 🟦2	🟥6 🟦4	🟥2 🟦1	🟥1 🟦1	🟦1	🟥2 🟦1	🟥1 🟦3	🟥1 🟦1		🟥6 🟦4	🟥1 🟦2		🟥2 🟦3	🟥1 🟦2	🟥2 🟦4	🟥5 🟦1	🟥3 🟦1
CD44	🟥25 🟦25	🟥8 🟦6	🟦1	🟥4 🟦2	🟥4 🟦6	🟥2	🟥1		🟥1 🟦1	🟦3			🟥4 🟦8	🟥3 🟦1		🟥2	🟥1 🟦2	🟥2 🟦4	🟥4 🟦1	🟥3 🟦1
JUP	🟥22 🟦23	🟥7 🟦7	🟦1	🟥3 🟦1	🟥5 🟦7	🟥3	🟥1	🟥2	🟥2	🟥1 🟦2			🟥6 🟦2	🟦3		🟥2	🟥2 🟦4	🟥2 🟦1	🟥4 🟦1	🟥2 🟦1
MACF1	🟥17 🟦27	🟥4 🟦10	🟥1 🟦2	🟥4 🟦1	🟥2 🟦6	🟦3	🟥1	🟦1	🟥1 🟦1	🟥1 🟦2	🟥1		🟥2 🟦6	🟥1 🟦3	🟥1	🟥2 🟦1	🟥1 🟦2	🟥3 🟦2	🟥3	🟥3 🟦1
SOX4	🟥37 🟦9	🟥12 🟦3	🟥1	🟥4 🟦2	🟥5 🟦1	🟦2	🟥1	🟥1	🟥2	🟥2 🟦2			🟥8	🟥1 🟦1		🟥1	🟥3	🟥3 🟦1	🟥2 🟦2	🟥3 🟦1
PPM1A	🟥22 🟦18	🟥6 🟦6	🟥2	🟥1 🟦3	🟥5 🟦8	🟥1 🟦2	🟥1		🟦1	🟥2 🟦2		🟥1	🟥2 🟦4	🟥2 🟦3	🟥1	🟦1	🟥3	🟥3 🟦3	🟥1 🟦1	🟥1
DAB2	🟥15 🟦28	🟥4 🟦8	🟦1	🟥3 🟦1	🟥3 🟦8	🟥2		🟦1	🟥1	🟥2	🟥1		🟥2 🟦7	🟥3 🟦1		🟥1 🟦2	🟥1 🟦2	🟥3 🟦2	🟥1 🟦2	🟦1

Atlas query refining (method 2)

Genes ?

Wnt receptor signali… x

up/down in

Organism

Homo sapiens

Conditions ?

carcinoma x

View

● Heatmap

○ List

Search Atlas

advanced search

e.g. ASPM, "p53 binding"

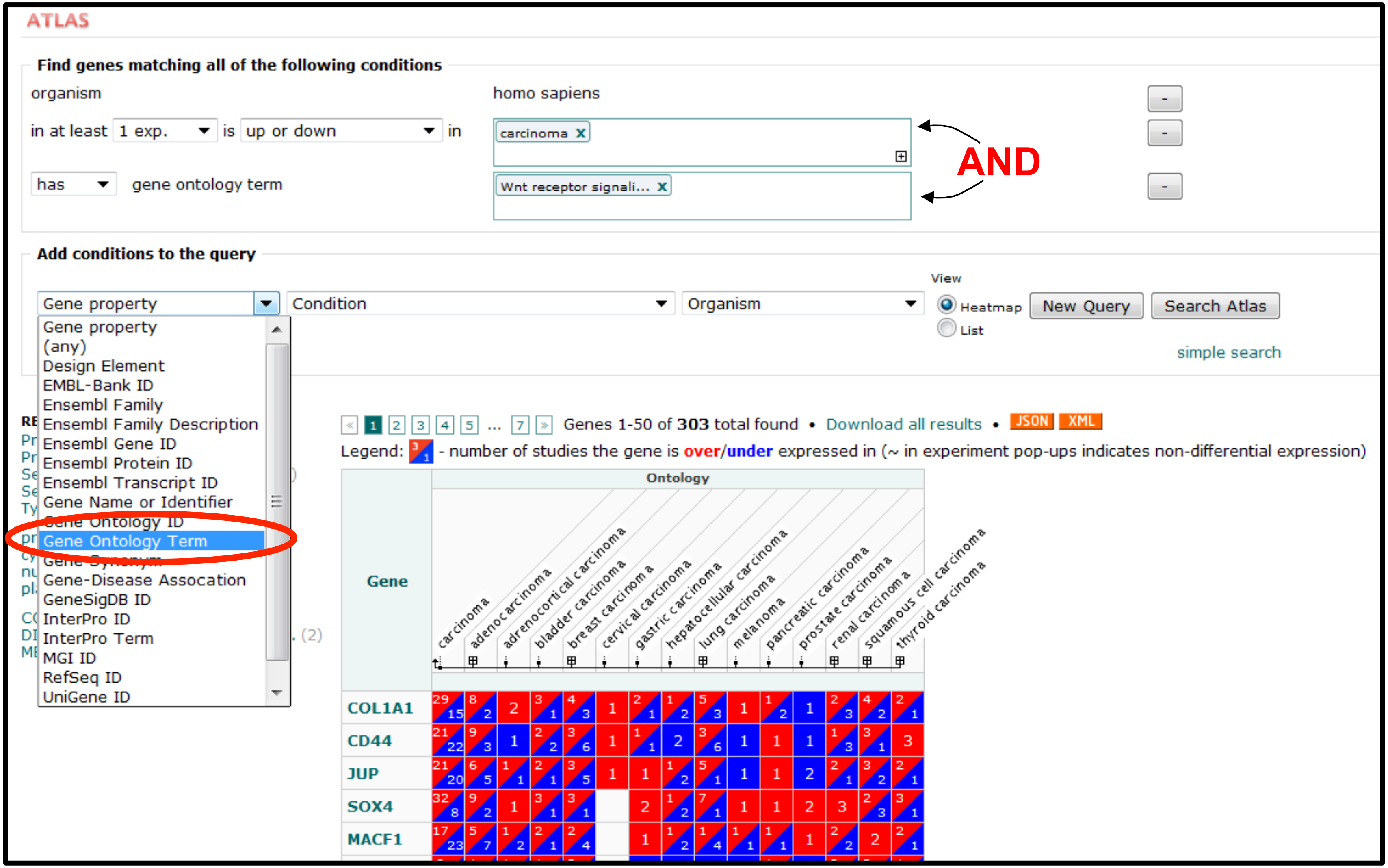
e.g. liver, cancer, diabetes

« 1 2 3 4 5 6 » Genes 1-50 of **288** total found (you can refine your query) • Download all results • JSON XML

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

Gene	carcinoma	adenocarcinoma	adrenocortical carcinoma	bladder carcinoma	breast carcinoma	cervical carcinoma	endometrioid carcinoma	esophageal carcinoma	gastric carcinoma	hepatocellular carcinoma	hereditary leiomyomatosis an...	large cell neuroendocrine car...	lung carcinoma	melanoma	ovarian carcinoma	pancreatic carcinoma	prostate carcinoma	renal carcinoma	squamous cell carcinoma	thyroid carcinoma
COL1A1	🟥34 🟦19	🟥9 🟦6	🟥2	🟥4 🟦2	🟥6 🟦4	🟥2 🟦1	🟥1 🟦1	🟦1	🟥2 🟦1	🟥1 🟦3	🟥1 🟦1		🟥6 🟦4	🟥1 🟦2		🟥2 🟦3	🟥1 🟦2	🟥2 🟦4	🟥5 🟦1	🟥3 🟦1
CD44	🟥25 🟦25	🟥8 🟦6	🟦1	🟥4 🟦2	🟥4 🟦6	🟥2	🟥1		🟥1 🟦1	🟦3			🟥4 🟦8	🟥3 🟦1		🟥2	🟥1 🟦2	🟥2 🟦4	🟥4 🟦1	🟥3 🟦1
JUP	🟥22 🟦23	🟥7 🟦7	🟦1	🟥3 🟦1	🟥5 🟦7	🟥3	🟥1	🟥2	🟥2	🟥1 🟦2			🟥6 🟦2	🟦3		🟥2	🟥2 🟦4	🟥2 🟦1	🟥4 🟦1	🟥2 🟦1
MACF1	🟥17 🟦27	🟥4 🟦10	🟥1 🟦2	🟥4 🟦1	🟥2 🟦6	🟦3	🟥1	🟦1	🟥1 🟦1	🟥1 🟦2	🟥1		🟥2 🟦6	🟥1 🟦3	🟥1	🟥2 🟦1	🟥1 🟦2	🟥3 🟦2	🟥3	🟥3 🟦1
SOX4	🟥37 🟦9	🟥12 🟦3	🟥1	🟥4 🟦2	🟥5 🟦1	🟦2	🟥1	🟥1	🟥2	🟥2 🟦2			🟥8	🟥1 🟦1		🟥1	🟥3	🟥3 🟦1	🟥2 🟦2	🟥3 🟦1
PPM1A	🟥22 🟦18	🟥6 🟦6	🟥2	🟥1 🟦3	🟥5 🟦8	🟥1 🟦2	🟥1		🟦1	🟥2 🟦2		🟥1	🟥2 🟦4	🟥2 🟦3	🟥1	🟦1	🟥3	🟥3 🟦3	🟥1 🟦1	🟥1
DAB2	🟥15 🟦28	🟥4 🟦8	🟦1	🟥3 🟦1	🟥3 🟦8	🟥2		🟦1	🟥1	🟥2	🟥1		🟥2 🟦7	🟥3 🟦1		🟥1 🟦2	🟥1 🟦2	🟥3 🟦2	🟥1 🟦2	🟦1

Atlas query refining (method 2)



ATLAS

Find genes matching all of the following conditions

organism: homo sapiens

in at least 1 exp. is up or down in: carcinoma x

has gene ontology term: Wnt receptor signali… x

AND

Add conditions to the query

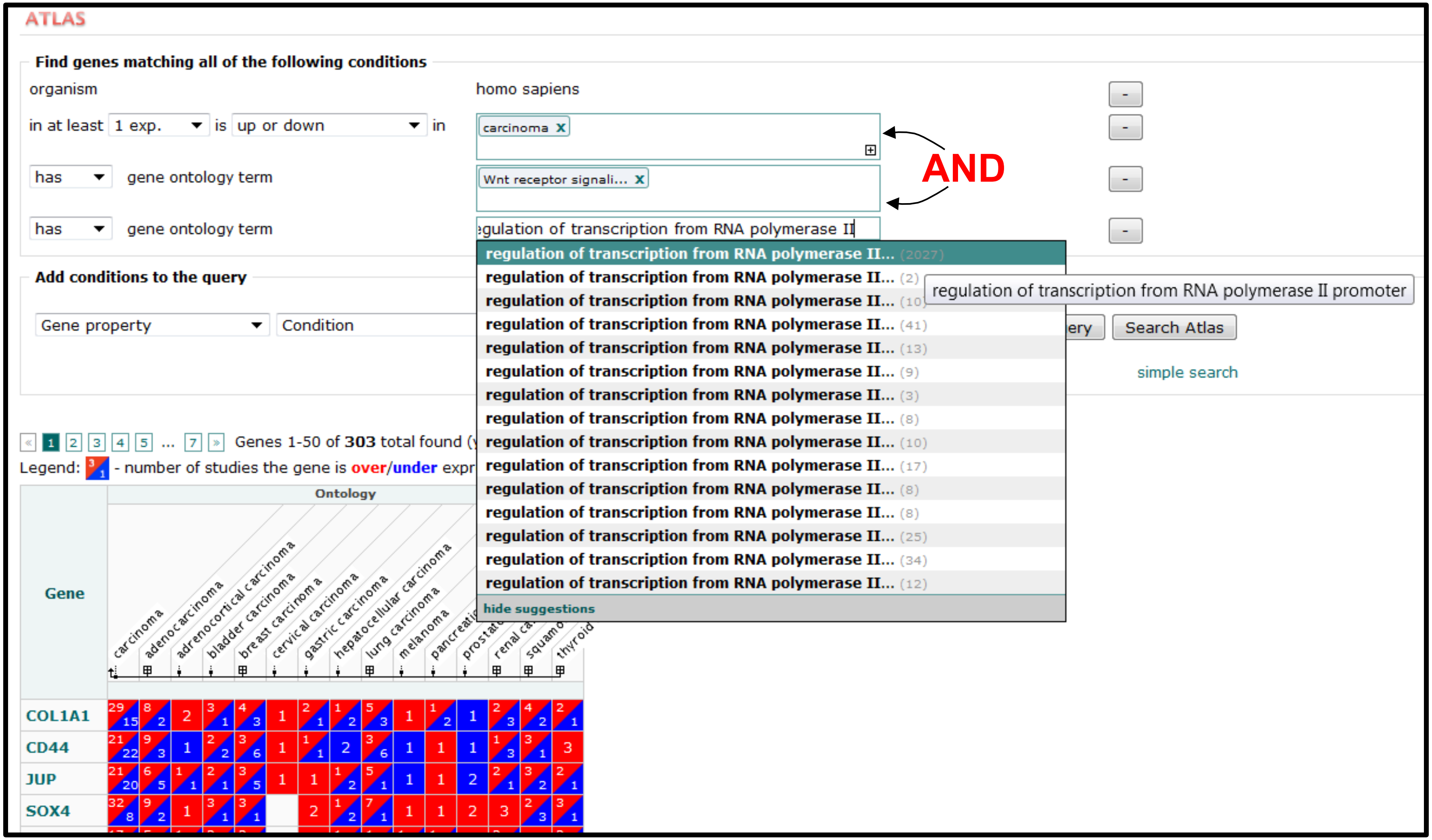
Genes 1-50 of **303** total found • Download all results •

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

Ontology

Gene	carcinoma	adenocarcinoma	adrenocortical carcinoma	bladder carcinoma	breast carcinoma	cervical carcinoma	gastric carcinoma	hepatocellular carcinoma	lung carcinoma	melanoma	pancreatic carcinoma	prostate carcinoma	renal carcinoma	squamous cell carcinoma	thyroid carcinoma
COL1A1	29 / 15	8 / 2	2	3 / 1	4 / 3	1	2 / 1	1 / 2	5 / 3	1	1 / 2	1	2 / 3	4 / 2	2 / 1
CD44	21 / 22	9 / 3	1	2 / 2	3 / 6	1	1 / 1	2	3 / 6	1	1	1	1 / 3	3 / 1	3
JUP	21 / 20	6 / 5	1 / 1	2 / 1	3 / 5	1	1	1 / 2	5 / 1	1	1	2	2 / 1	3 / 2	2 / 1
SOX4	32 / 8	9 / 2	1	3 / 1	3 / 1		2	1 / 2	7 / 1	1	1	2	3	2 / 3	3 / 1
MACF1	17 / 23	5 / 7	1 / 2	2 / 1	2 / 4		1	1 / 2	1 / 4	1 / 1	1 / 1	1	2 / 2	2	2 / 1

Atlas query refining (method 2)



ATLAS

Find genes matching all of the following conditions

organism — homo sapiens

in at least 1 exp. is up or down in — carcinoma x

has gene ontology term — Wnt receptor signali... x

has gene ontology term — egulation of transcription from RNA polymerase II

regulation of transcription from RNA polymerase II... (2027)
regulation of transcription from RNA polymerase II... (2)
regulation of transcription from RNA polymerase II... (10)
regulation of transcription from RNA polymerase II... (41)
regulation of transcription from RNA polymerase II... (13)
regulation of transcription from RNA polymerase II... (9)
regulation of transcription from RNA polymerase II... (3)
regulation of transcription from RNA polymerase II... (8)
regulation of transcription from RNA polymerase II... (10)
regulation of transcription from RNA polymerase II... (17)
regulation of transcription from RNA polymerase II... (8)
regulation of transcription from RNA polymerase II... (8)
regulation of transcription from RNA polymerase II... (25)
regulation of transcription from RNA polymerase II... (34)
regulation of transcription from RNA polymerase II... (12)

regulation of transcription from RNA polymerase II promoter

Add conditions to the query

Gene property — Condition

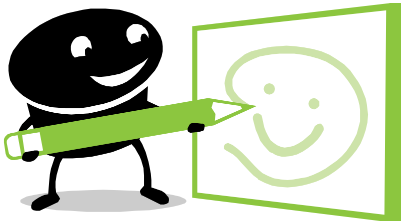
Genes 1-50 of **303** total found (

Legend: 🟥3 1 - number of studies the gene is **over**/**under** expr

Gene	carcinoma	adenocarcinoma	adrenocortical carcinoma	bladder carcinoma	breast carcinoma	cervical carcinoma	gastric carcinoma	hepatocellular carcinoma	lung carcinoma	melanoma	pancreati	prostat	renal ca	squamo	thyroid
COL1A1	29 15	8 2	2	3 1	4 3	1	2 1	1 2	5 3	1	1 2	1	2 3	4 2	2 1
CD44	21 22	9 3	1	2 2	3 6	1	1 1	2	3 6	1	1	1	1 3	3 1	3
JUP	21 20	6 5	1 1	2 1	3 5	1	1	1 2	5 1	1	1	2	2 1	3 2	2 1
SOX4	32 8	9 2	1	3 1	3 1		2	1 2	7 1	1	1	2	3	2 3	3 1

Hands-on exercise 3

Find genes in the “androgen receptor signaling pathway” which are (i) expressed in prostate carcinoma and (ii) involved in regulation of transcription from RNA Pol II



Hands-on exercise 4

Find information on *Tbx5* expression in mouse in relation to Holt-Oram syndrome

NEW

Expression Atlas

which genes are expressed in a normal human kidney?

“baseline”

which genes are up-regulated in pancreatic islets of pregnant mice?

“differential”

RNA-Seq data in new Atlas

metadata

ArrayExpress

ENA

FASTQ files

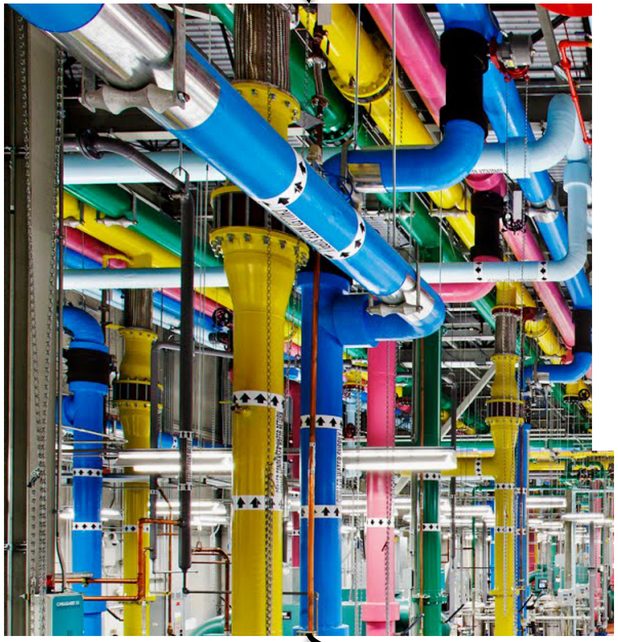
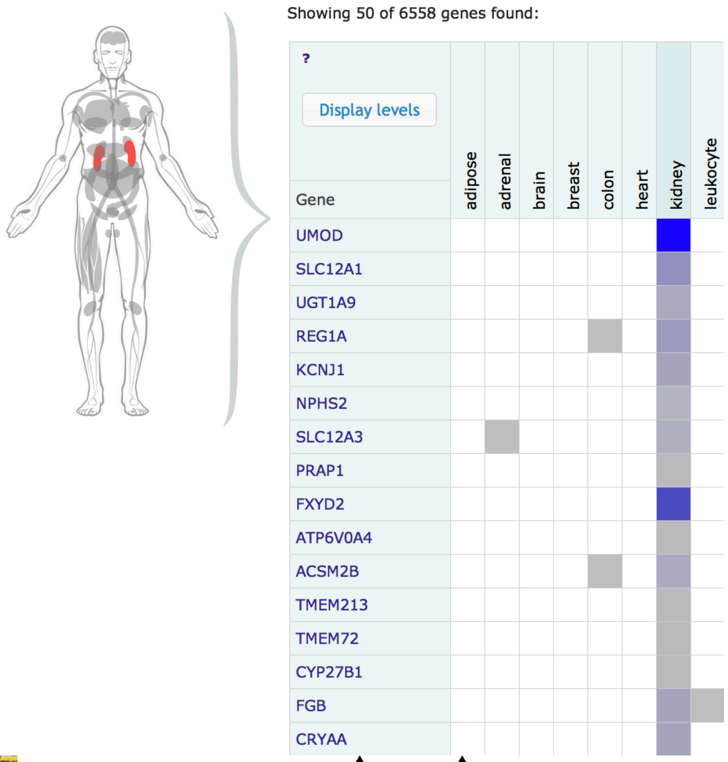


Photo © Google/Connie Zhou

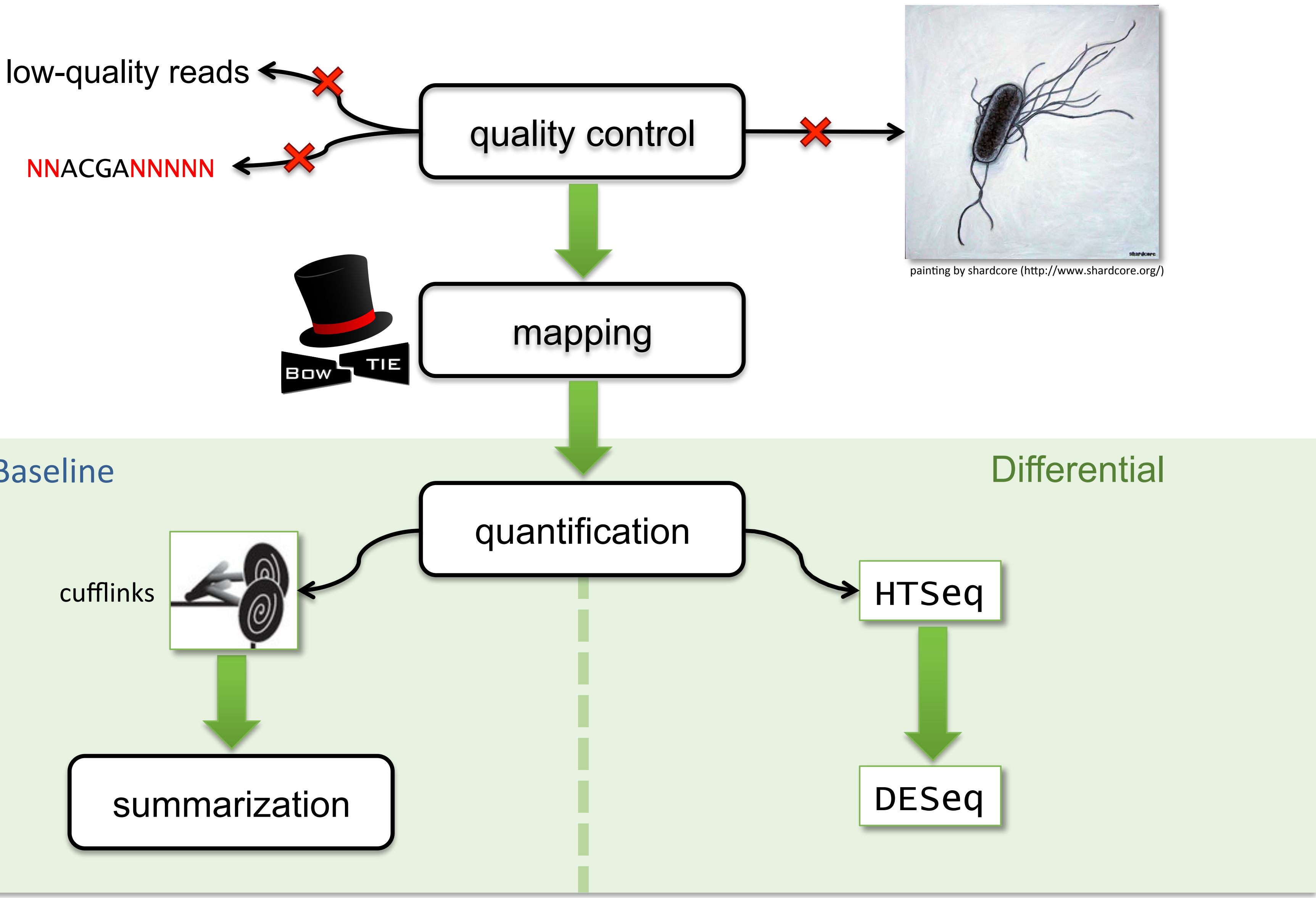


Showing 50 of 6558 genes found:

Gene	adipose	adrenal	brain	breast	colon	heart	kidney	leukocyte
UMOD							■	
SLC12A1							■	
UGT1A9							■	
REG1A					■		■	
KCNJ1							■	
NPHS2							■	
SLC12A3		■					■	
PRAP1							■	
FXYD2							■	
ATP6V0A4							■	
ACSM2B					■		■	
TMEM213							■	
TMEM72							■	
CYP27B1							■	
FGB							■	■
CRYAA							■	

Atlas results

New Atlas RNA-Seq pipeline



painting by shardcore (http://www.shardcore.org/)

Fonseca, N.A. et al (2013) iRAP – an integrated RNA-seq Analysis Pipeline, *Bioinformatics*, submitted

Expression Atlas

NEW

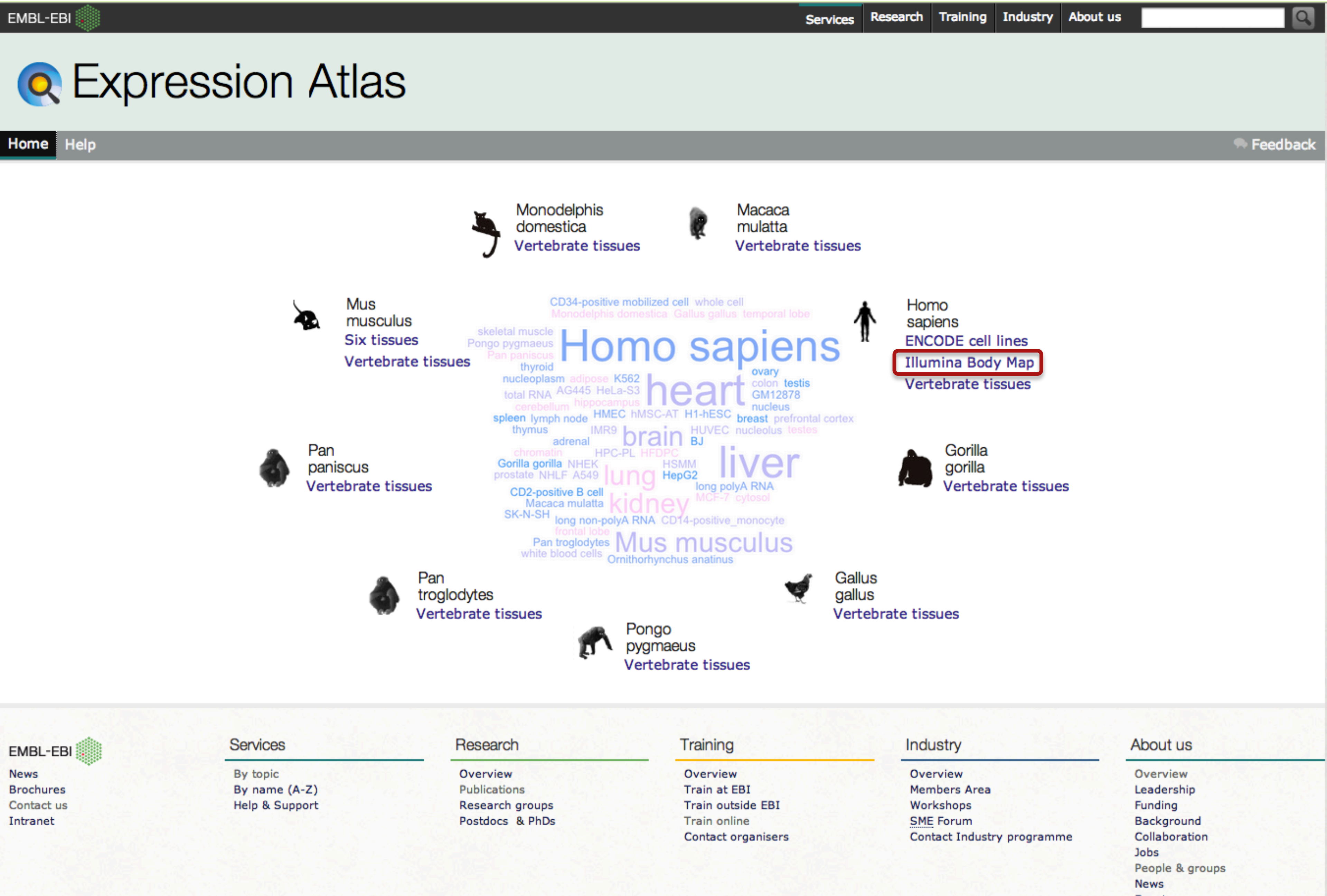
which genes are expressed in a normal human kidney?

"baseline"

which genes are up-regulated in pancreatic islets of pregnant mice?

"differential"

www-test.ebi.ac.uk/gxa



Expression Atlas

Monodelphis domestica
Vertebrate tissues

Macaca mulatta
Vertebrate tissues

Mus musculus
Six tissues
Vertebrate tissues

Homo sapiens
ENCODE cell lines
Illumina Body Map
Vertebrate tissues

Pan paniscus
Vertebrate tissues

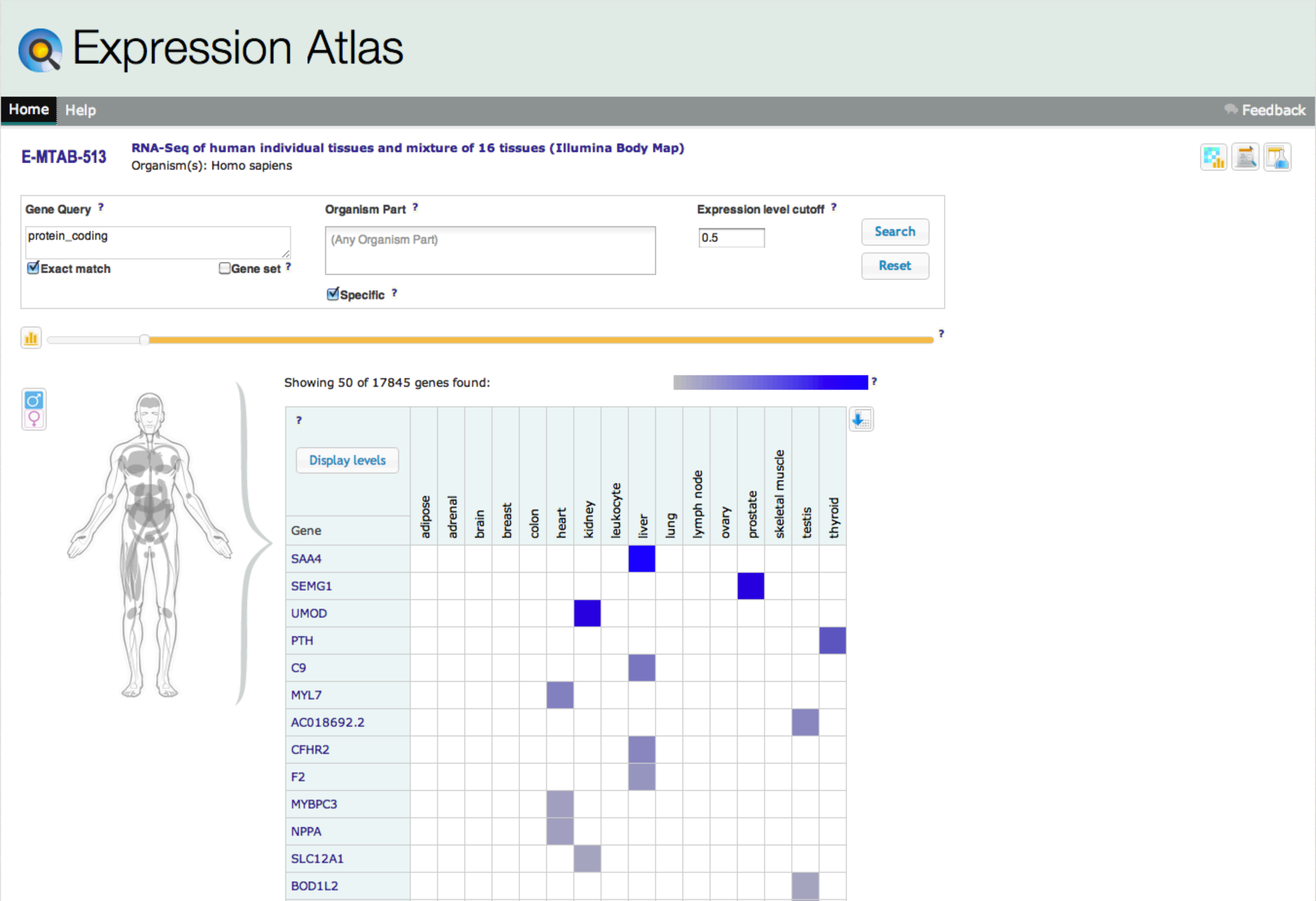
Gorilla gorilla
Vertebrate tissues

Pan troglodytes
Vertebrate tissues

Gallus gallus
Vertebrate tissues

Pongo pygmaeus
Vertebrate tissues

http://www-test.ebi.ac.uk/gxa/experiments/E-MTAB-513



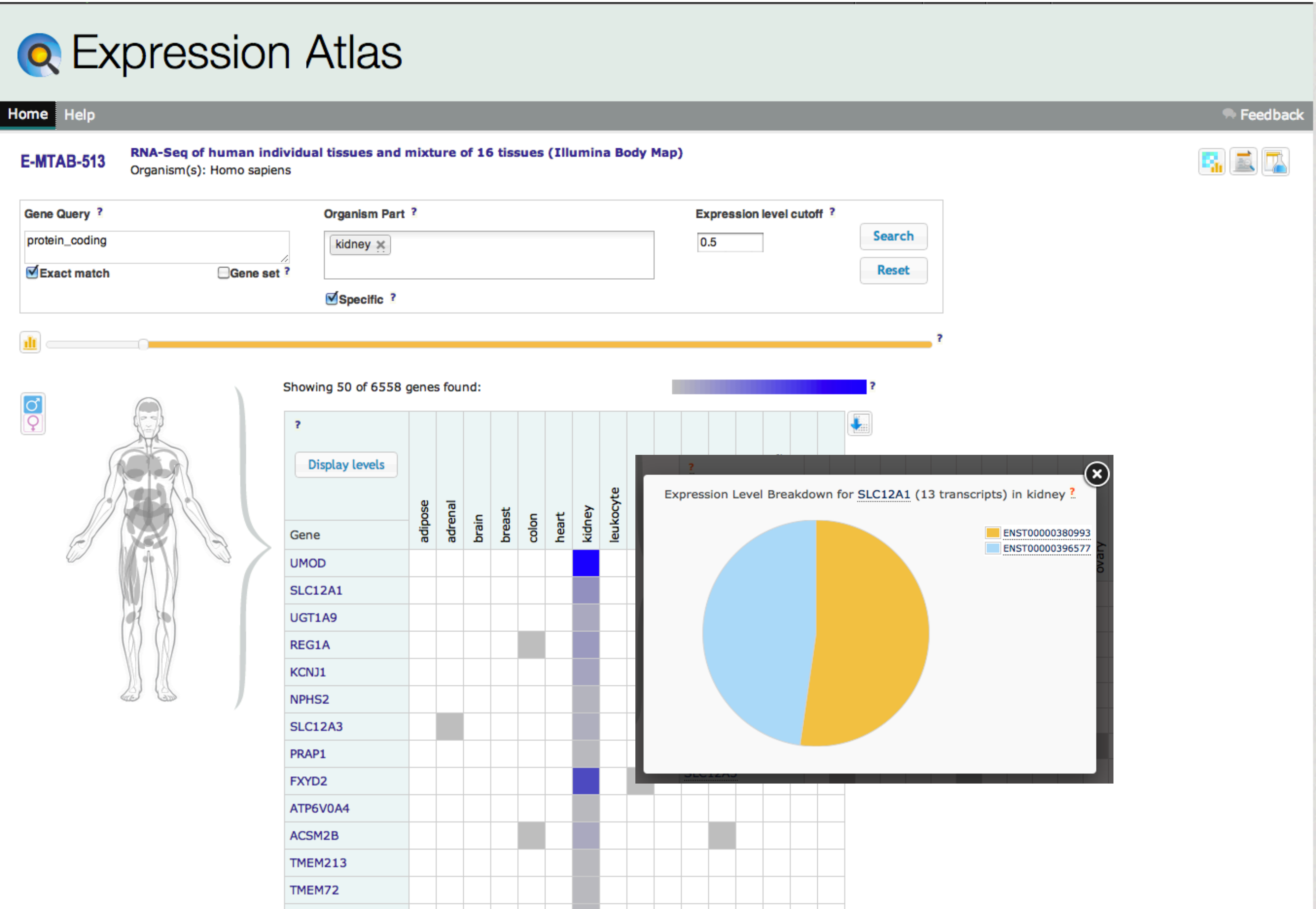
E-MTAB-513 **RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)**
Organism(s): Homo sapiens

Gene Query: protein_coding
☑ Exact match ☐ Gene set
Organism Part: (Any Organism Part)
☑ Specific
Expression level cutoff: 0.5

Showing 50 of 17845 genes found:

Gene	adipose	adrenal	brain	breast	colon	heart	kidney	leukocyte	liver	lung	lymph node	ovary	prostate	skeletal muscle	testis	thyroid
SAA4									■							
SEMG1													■			
UMOD							■									
PTH																■
C9									■							
MYL7						■										
AC018692.2															■	
CFHR2									■							
F2									■							
MYBPC3						■										
NPPA						■										
SLC12A1							■									
BOD1L2															■	

http://www-test.ebi.ac.uk/gxa/experiments/E-MTAB-513



E-MTAB-513 RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)

Organism(s): Homo sapiens

Gene Query: protein_coding

☑ Exact match ☐ Gene set

Organism Part: kidney

☑ Specific

Expression level cutoff: 0.5

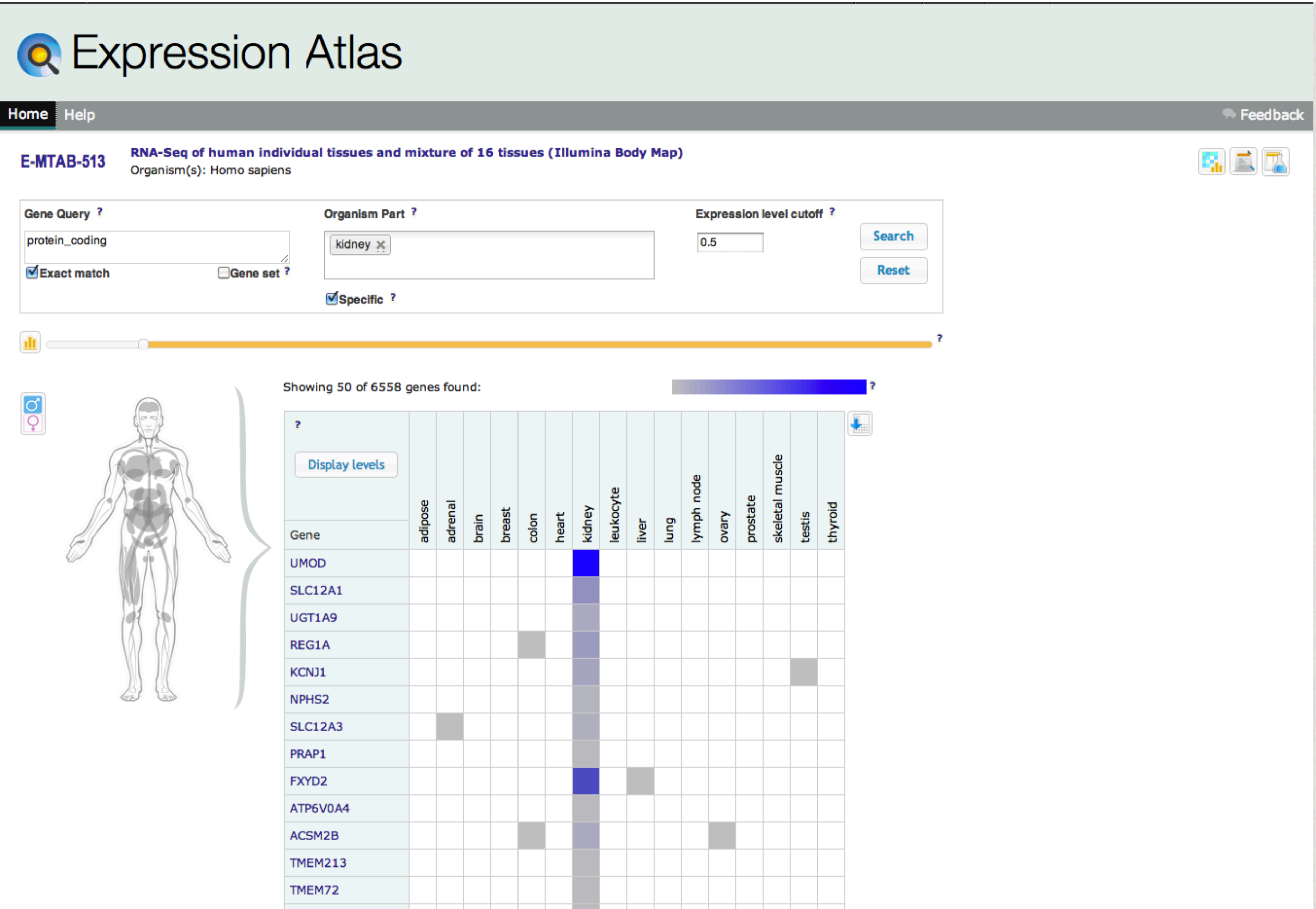
Search

Reset

Showing 50 of 6558 genes found:

Expression Level Breakdown for SLC12A1 (13 transcripts) in kidney

http://www-test.ebi.ac.uk/gxa/experiments/E-MTAB-513



E-MTAB-513 **RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)**
Organism(s): Homo sapiens

Gene Query ? protein_coding
☑ Exact match ☐ Gene set ?

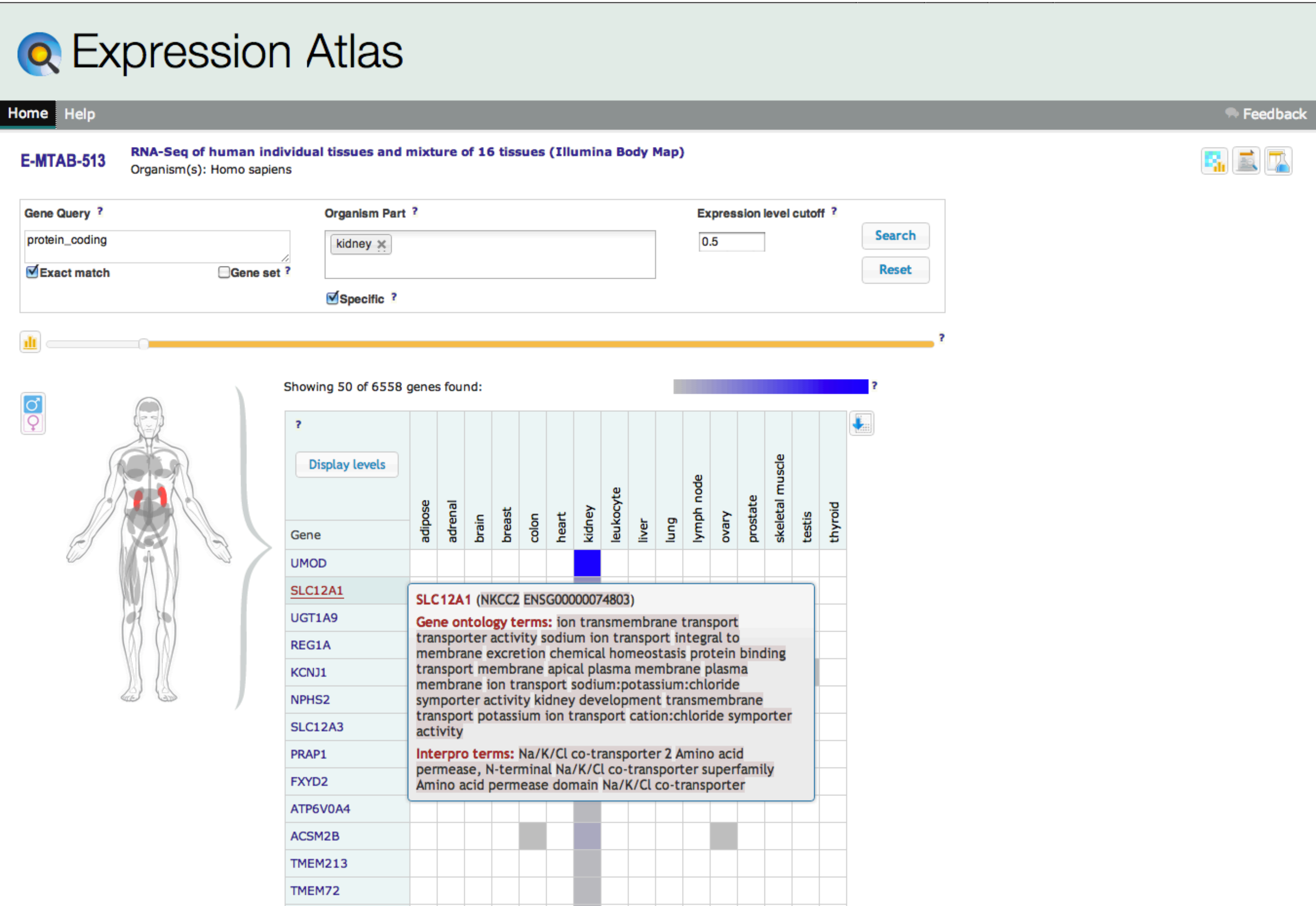
Organism Part ? kidney ✕
☑ Specific ?

Expression level cutoff ? 0.5

Showing 50 of 6558 genes found:

Gene	adipose	adrenal	brain	breast	colon	heart	kidney	leukocyte	liver	lung	lymph node	ovary	prostate	skeletal muscle	testis	thyroid
UMOD							■									
SLC12A1							■									
UGT1A9							■									
REG1A					■		■									
KCNJ1							■								■	
NPHS2							■									
SLC12A3		■					■									
PRAP1							■									
FXYD2							■		■							
ATP6V0A4							■									
ACSM2B					■		■					■				
TMEM213							■									
TMEM72							■									

http://www-test.ebi.ac.uk/gxa/experiments/E-MTAB-513



E-MTAB-513 **RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)**
Organism(s): Homo sapiens

Gene Query ?: protein_coding
☑ Exact match ☐ Gene set ?

Organism Part ?: kidney ×
☑ Specific ?

Expression level cutoff ?: 0.5

Showing 50 of 6558 genes found:

SLC12A1 (NKCC2 ENSG00000074803)

Gene ontology terms: ion transmembrane transport transporter activity sodium ion transport integral to membrane excretion chemical homeostasis protein binding transport membrane apical plasma membrane plasma membrane ion transport sodium:potassium:chloride symporter activity kidney development transmembrane transport potassium ion transport cation:chloride symporter activity

Interpro terms: Na/K/Cl co-transporter 2 Amino acid permease, N-terminal Na/K/Cl co-transporter superfamily Amino acid permease domain Na/K/Cl co-transporter

Expression Atlas

NEW

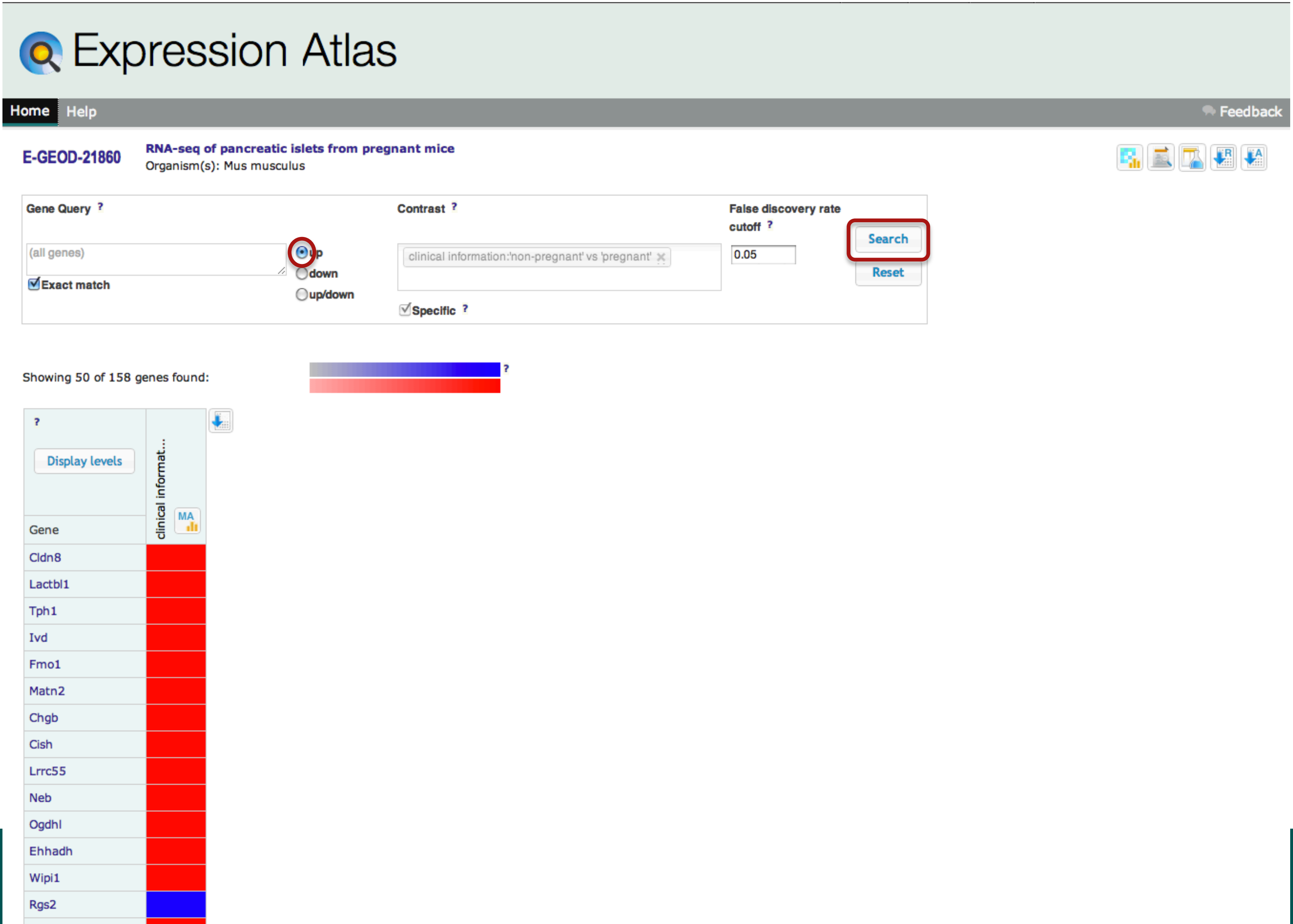
which genes are expressed in a normal human kidney?

"baseline"

which genes are up-regulated in pancreatic islets of pregnant mice?

"differential"

http://www-test.ebi.ac.uk/gxa/experiments/E-GEOD-21860



E-GEOD-21860 **RNA-seq of pancreatic islets from pregnant mice**
Organism(s): Mus musculus

Gene Query ? (all genes)
☑ Exact match
● up
○ down
○ up/down

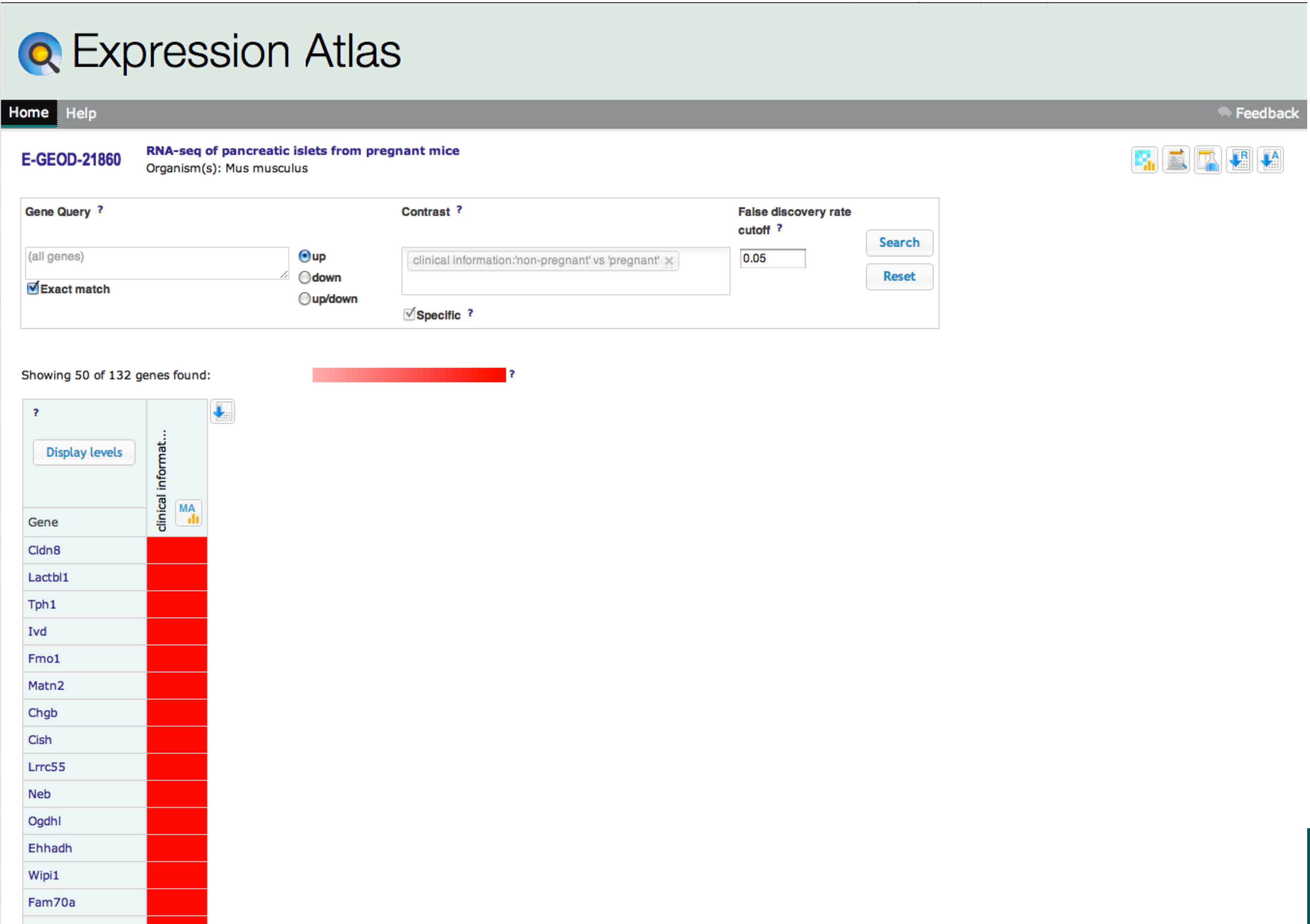
Contrast ? clinical information:'non-pregnant' vs 'pregnant'
☑ Specific ?

False discovery rate cutoff ? 0.05

Showing 50 of 158 genes found:

Gene	clinical informat...
Cldn8	🟥
Lactbl1	🟥
Tph1	🟥
Ivd	🟥
Fmo1	🟥
Matn2	🟥
Chgb	🟥
Cish	🟥
Lrrc55	🟥
Neb	🟥
Ogdhl	🟥
Ehhadh	🟥
Wipi1	🟥
Rgs2	🟦

http://www-test.ebi.ac.uk/gxa/experiments/E-GEOD-21860



E-GEOD-21860 RNA-seq of pancreatic islets from pregnant mice

Organism(s): Mus musculus

Gene Query ?: (all genes)

☑ Exact match

● up
○ down
○ up/down

Contrast ?: clinical information:'non-pregnant' vs 'pregnant'

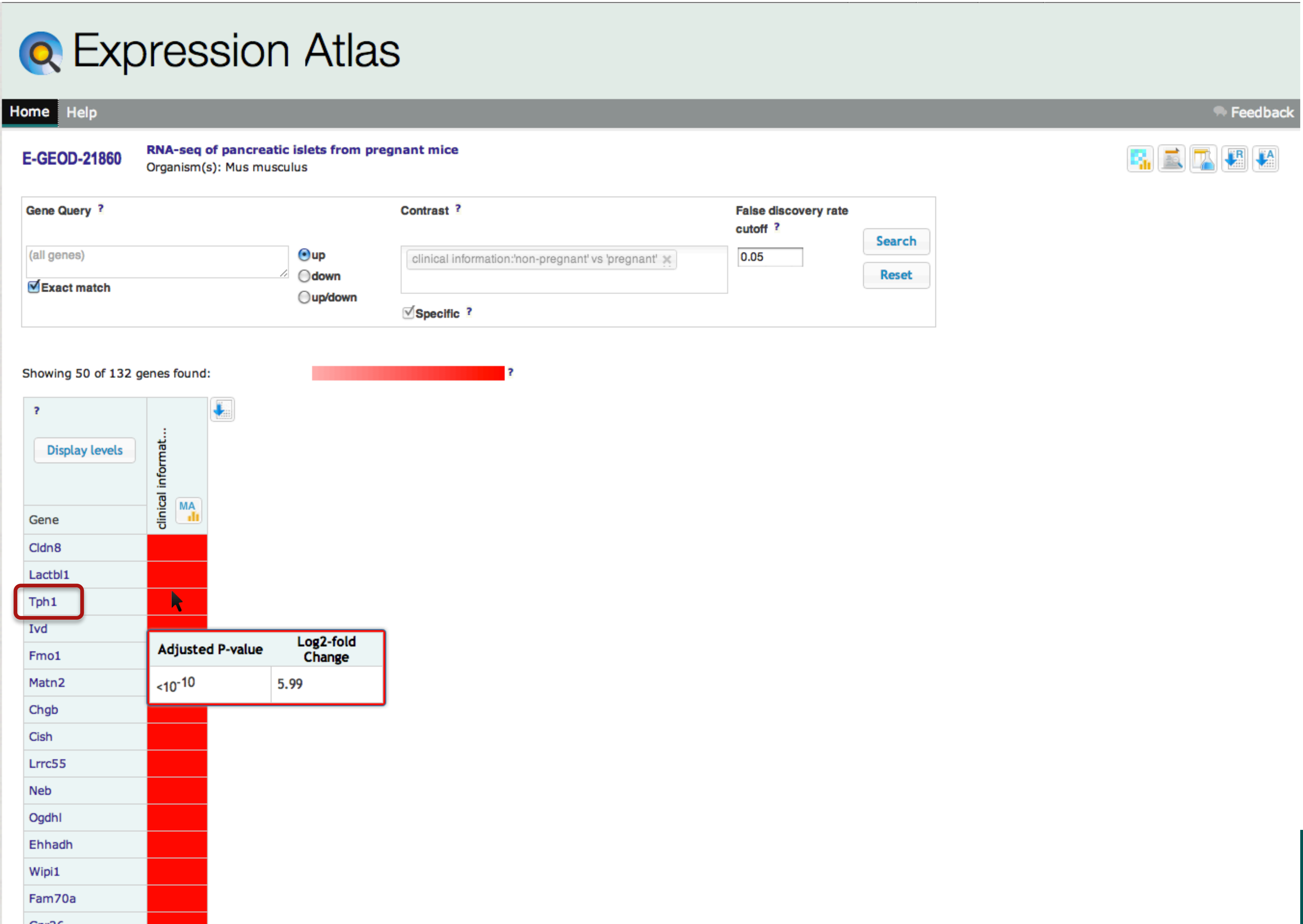
☑ Specific ?

False discovery rate cutoff ?: 0.05

Showing 50 of 132 genes found:

Gene	clinical informat...
Cldn8	🟥
Lactbl1	🟥
Tph1	🟥
Ivd	🟥
Fmo1	🟥
Matn2	🟥
Chgb	🟥
Cish	🟥
Lrrc55	🟥
Neb	🟥
Ogdhl	🟥
Ehhadh	🟥
Wipi1	🟥
Fam70a	🟥

http://www-test.ebi.ac.uk/gxa/experiments/E-GEOD-21860



E-GEOD-21860 **RNA-seq of pancreatic islets from pregnant mice**
Organism(s): Mus musculus

Gene Query ? (all genes)
☑ Exact match
● up
○ down
○ up/down

Contrast ? clinical information:'non-pregnant' vs 'pregnant'
☑ Specific ?

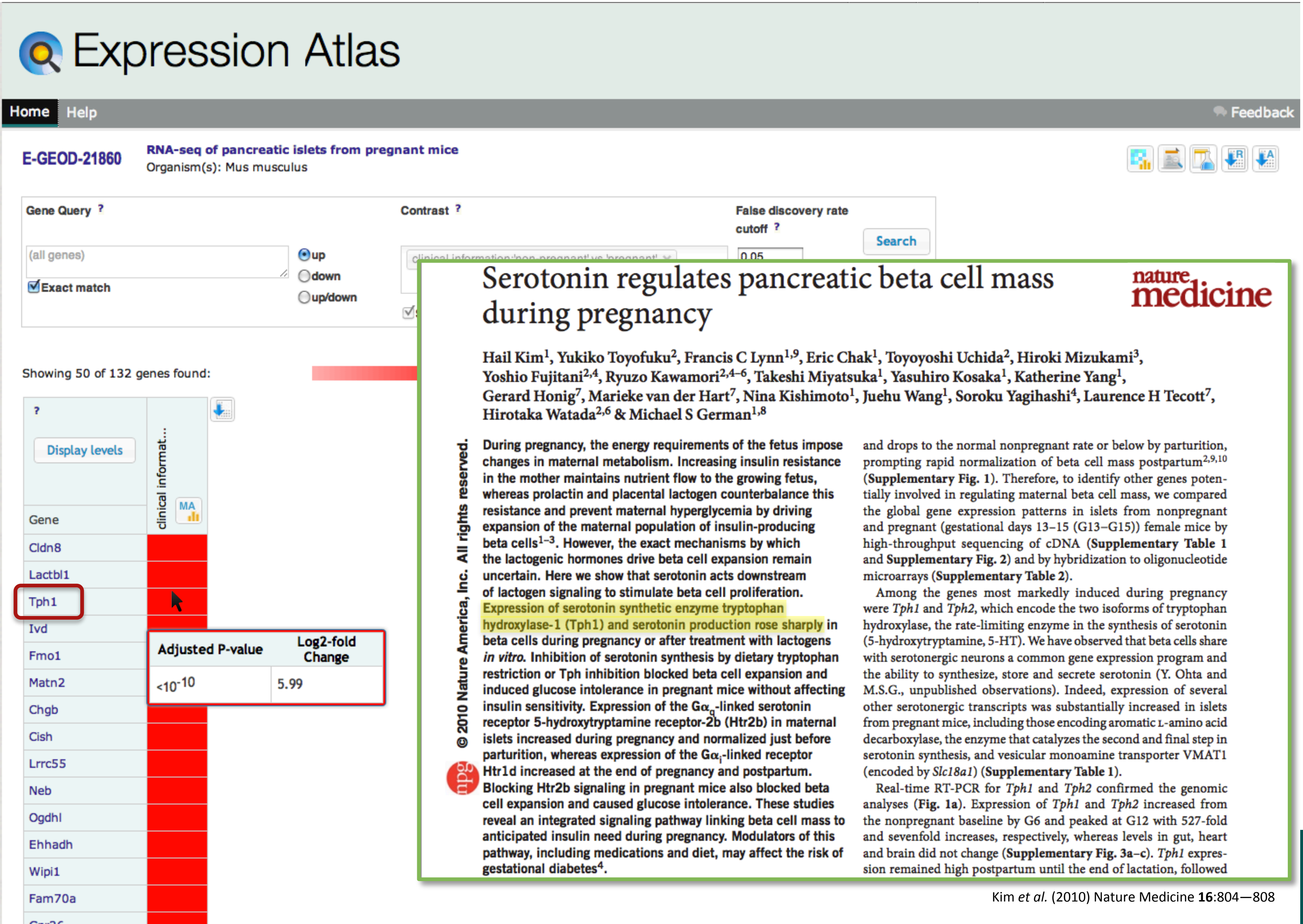
False discovery rate cutoff ? 0.05

Showing 50 of 132 genes found:

Gene	clinical informat...
Cldn8	🟥
Lactbl1	🟥
Tph1	🟥
Ivd	🟥
Fmo1	🟥
Matn2	🟥
Chgb	🟥
Cish	🟥
Lrrc55	🟥
Neb	🟥
Ogdhl	🟥
Ehhadh	🟥
Wipi1	🟥
Fam70a	🟥

Adjusted P-value	Log2-fold Change
$<10^{-10}$	5.99

http://www-test.ebi.ac.uk/gxa/experiments/E-GEOD-21860



E-GEOD-21860 **RNA-seq of pancreatic islets from pregnant mice**
Organism(s): Mus musculus

Gene Query ?
(all genes)
● up ○ down ○ up/down
☑ Exact match

Contrast ?
clinical information:'non-pregnant' vs 'pregnant'

False discovery rate cutoff ?
0.05

Search

Showing 50 of 132 genes found:

Gene
Cldn8
Lactbl1
Tph1
Ivd
Fmo1
Matn2
Chgb
Cish
Lrrc55
Neb
Ogdhl
Ehhadh
Wipi1
Fam70a

Adjusted P-value	Log2-fold Change
$<10^{-10}$	5.99

Serotonin regulates pancreatic beta cell mass during pregnancy

nature medicine

Hail Kim[1], Yukiko Toyofuku[2], Francis C Lynn[1,9], Eric Chak[1], Toyoyoshi Uchida[2], Hiroki Mizukami[3], Yoshio Fujitani[2,4], Ryuzo Kawamori[2,4–6], Takeshi Miyatsuka[1], Yasuhiro Kosaka[1], Katherine Yang[1], Gerard Honig[7], Marieke van der Hart[7], Nina Kishimoto[1], Juehu Wang[1], Soroku Yagihashi[4], Laurence H Tecott[7], Hirotaka Watada[2,6] & Michael S German[1,8]

© 2010 Nature America, Inc. All rights reserved.

During pregnancy, the energy requirements of the fetus impose changes in maternal metabolism. Increasing insulin resistance in the mother maintains nutrient flow to the growing fetus, whereas prolactin and placental lactogen counterbalance this resistance and prevent maternal hyperglycemia by driving expansion of the maternal population of insulin-producing beta cells[1–3]. However, the exact mechanisms by which the lactogenic hormones drive beta cell expansion remain uncertain. Here we show that serotonin acts downstream of lactogen signaling to stimulate beta cell proliferation. Expression of serotonin synthetic enzyme tryptophan hydroxylase-1 (Tph1) and serotonin production rose sharply in beta cells during pregnancy or after treatment with lactogens *in vitro*. Inhibition of serotonin synthesis by dietary tryptophan restriction or Tph inhibition blocked beta cell expansion and induced glucose intolerance in pregnant mice without affecting insulin sensitivity. Expression of the $G\alpha_q$-linked serotonin receptor 5-hydroxytryptamine receptor-2b (Htr2b) in maternal islets increased during pregnancy and normalized just before parturition, whereas expression of the $G\alpha_i$-linked receptor Htr1d increased at the end of pregnancy and postpartum. Blocking Htr2b signaling in pregnant mice also blocked beta cell expansion and caused glucose intolerance. These studies reveal an integrated signaling pathway linking beta cell mass to anticipated insulin need during pregnancy. Modulators of this pathway, including medications and diet, may affect the risk of gestational diabetes[4].

and drops to the normal nonpregnant rate or below by parturition, prompting rapid normalization of beta cell mass postpartum[2,9,10] (**Supplementary Fig. 1**). Therefore, to identify other genes potentially involved in regulating maternal beta cell mass, we compared the global gene expression patterns in islets from nonpregnant and pregnant (gestational days 13–15 (G13–G15)) female mice by high-throughput sequencing of cDNA (**Supplementary Table 1** and **Supplementary Fig. 2**) and by hybridization to oligonucleotide microarrays (**Supplementary Table 2**).

Among the genes most markedly induced during pregnancy were *Tph1* and *Tph2*, which encode the two isoforms of tryptophan hydroxylase, the rate-limiting enzyme in the synthesis of serotonin (5-hydroxytryptamine, 5-HT). We have observed that beta cells share with serotonergic neurons a common gene expression program and the ability to synthesize, store and secrete serotonin (Y. Ohta and M.S.G., unpublished observations). Indeed, expression of several other serotonergic transcripts was substantially increased in islets from pregnant mice, including those encoding aromatic L-amino acid decarboxylase, the enzyme that catalyzes the second and final step in serotonin synthesis, and vesicular monoamine transporter VMAT1 (encoded by *Slc18a1*) (**Supplementary Table 1**).

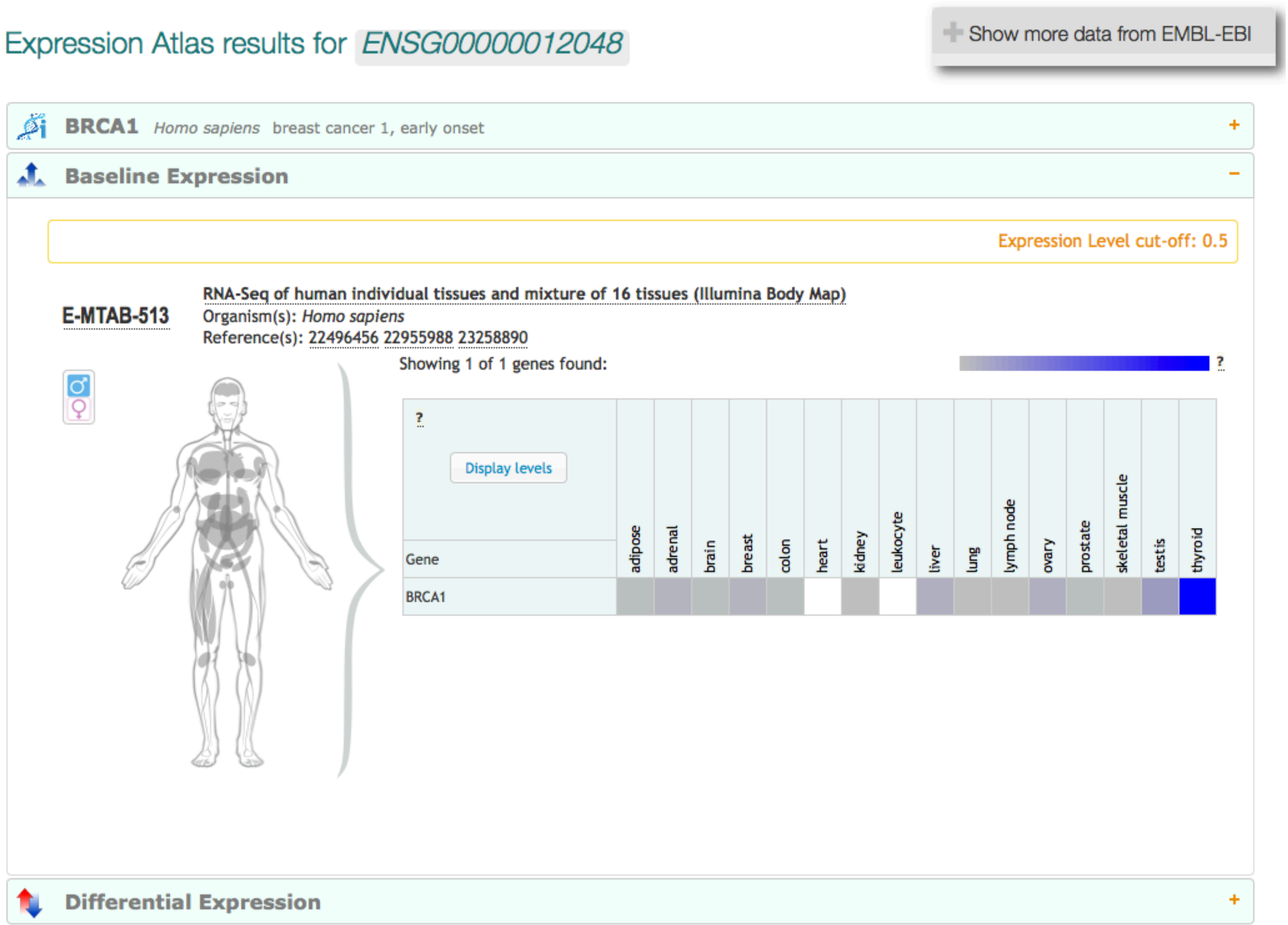
Real-time RT-PCR for *Tph1* and *Tph2* confirmed the genomic analyses (**Fig. 1a**). Expression of *Tph1* and *Tph2* increased from the nonpregnant baseline by G6 and peaked at G12 with 527-fold and sevenfold increases, respectively, whereas levels in gut, heart and brain did not change (**Supplementary Fig. 3a–c**). *Tph1* expression remained high postpartum until the end of lactation, followed

Kim *et al.* (2010) Nature Medicine **16**:804—808

Future plans for new Atlas

- Search across datasets
  - for genes
  - for conditions
- Gene sets and pathways
- Proteomics

Baseline Expression on page example for human BRCA1 gene: http://www-test.ebi.ac.uk/gxa/genes/ENSG00000012048



Expression Atlas results for *ENSG00000012048*

BRCA1 *Homo sapiens* breast cancer 1, early onset

Baseline Expression

Expression Level cut-off: 0.5

E-MTAB-513

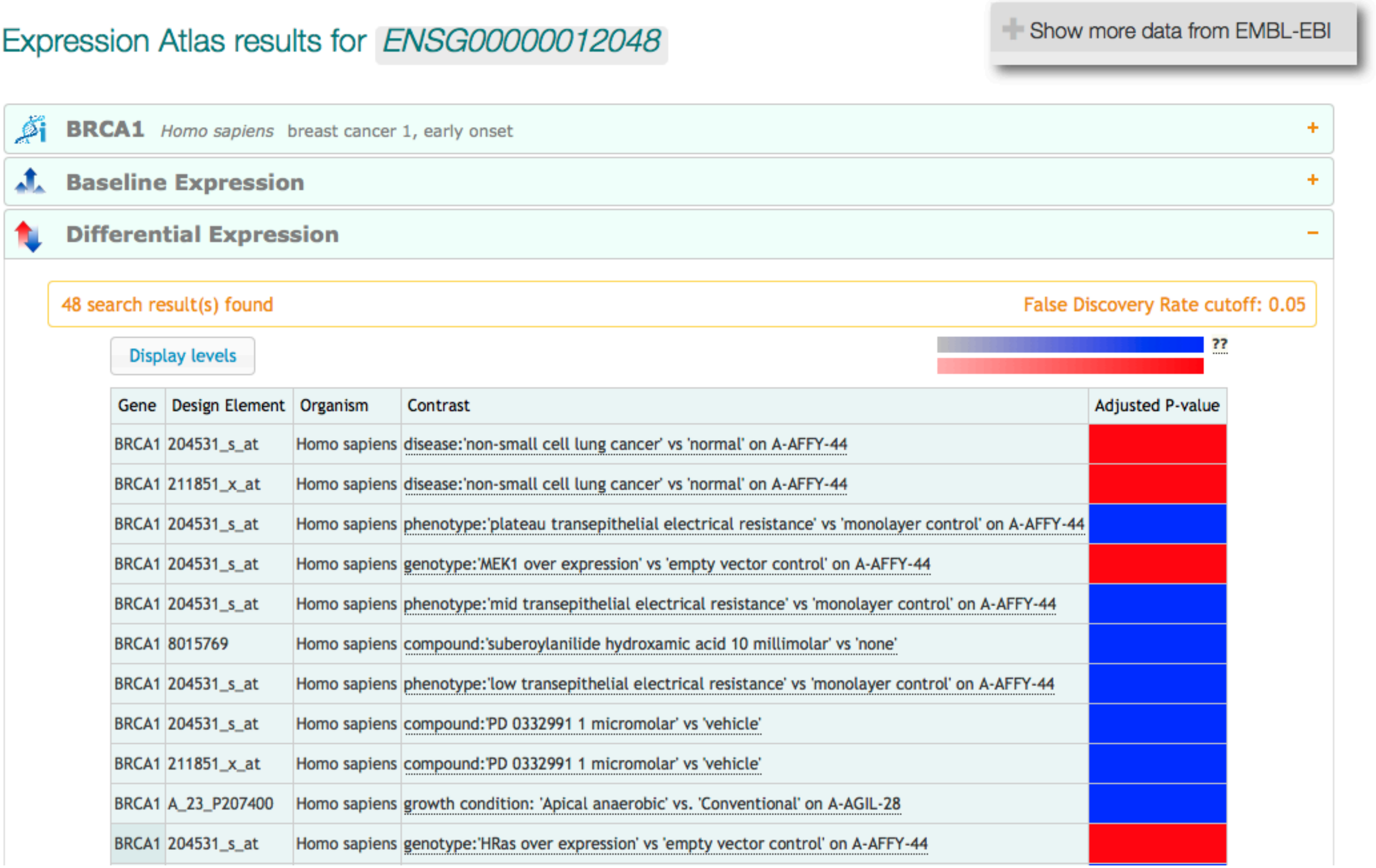
RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)
Organism(s): *Homo sapiens*
Reference(s): 22496456 22955988 23258890

Showing 1 of 1 genes found:

Gene	adipose	adrenal	brain	breast	colon	heart	kidney	leukocyte	liver	lung	lymph node	ovary	prostate	skeletal muscle	testis	thyroid
BRCA1																

Differential Expression

Baseline Expression on page example for human BRCA1 gene: http://www-test.ebi.ac.uk/gxa/genes/ENSG00000012048



Expression Atlas results for *ENSG00000012048*

BRCA1 *Homo sapiens* breast cancer 1, early onset

Baseline Expression

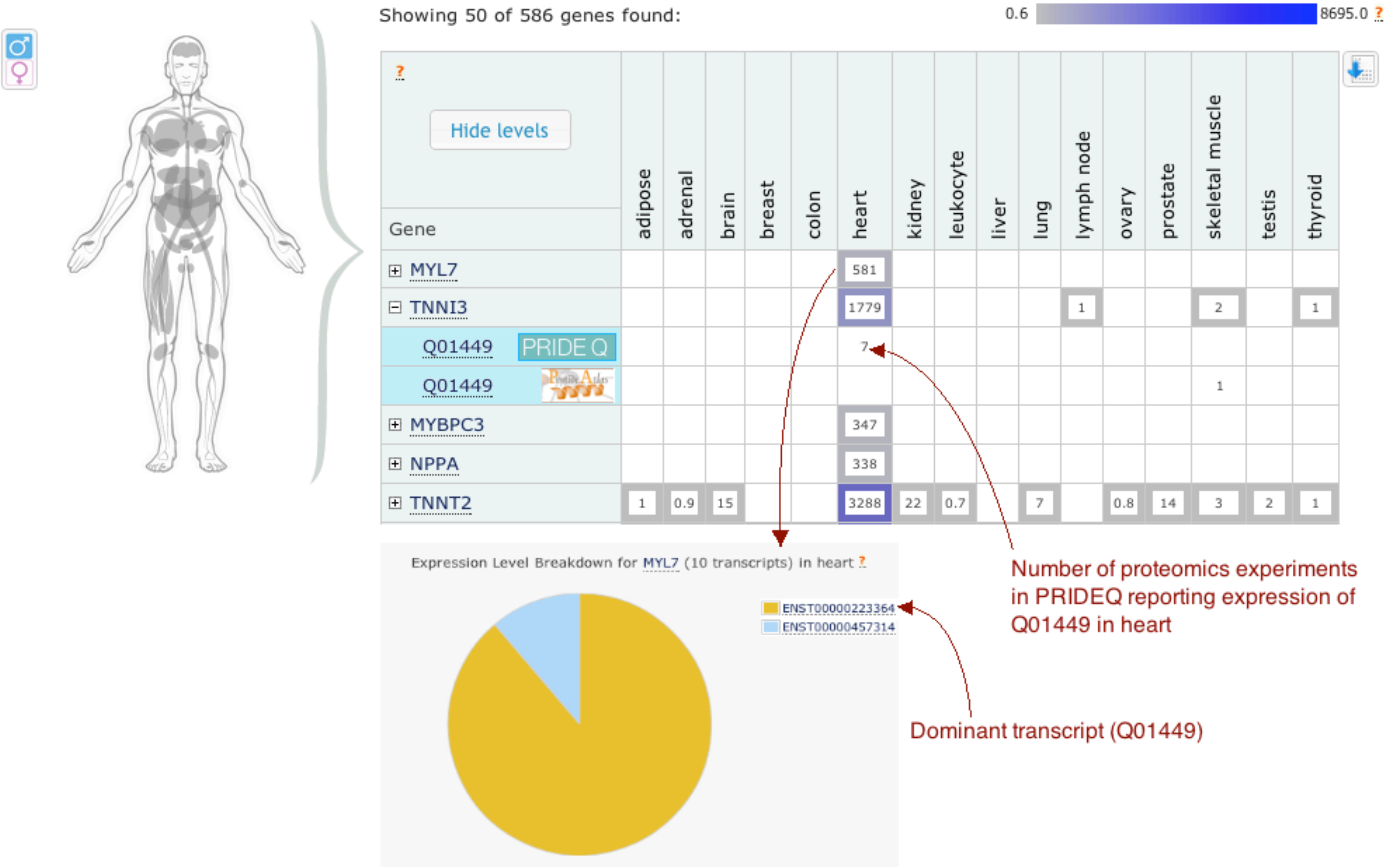
Differential Expression

48 search result(s) found

False Discovery Rate cutoff: 0.05

Gene	Design Element	Organism	Contrast	Adjusted P-value
BRCA1	204531_s_at	Homo sapiens	disease:'non-small cell lung cancer' vs 'normal' on A-AFFY-44	🟥
BRCA1	211851_x_at	Homo sapiens	disease:'non-small cell lung cancer' vs 'normal' on A-AFFY-44	🟥
BRCA1	204531_s_at	Homo sapiens	phenotype:'plateau transepithelial electrical resistance' vs 'monolayer control' on A-AFFY-44	🟦
BRCA1	204531_s_at	Homo sapiens	genotype:'MEK1 over expression' vs 'empty vector control' on A-AFFY-44	🟥
BRCA1	204531_s_at	Homo sapiens	phenotype:'mid transepithelial electrical resistance' vs 'monolayer control' on A-AFFY-44	🟦
BRCA1	8015769	Homo sapiens	compound:'suberoylanilide hydroxamic acid 10 millimolar' vs 'none'	🟦
BRCA1	204531_s_at	Homo sapiens	phenotype:'low transepithelial electrical resistance' vs 'monolayer control' on A-AFFY-44	🟦
BRCA1	204531_s_at	Homo sapiens	compound:'PD 0332991 1 micromolar' vs 'vehicle'	🟦
BRCA1	211851_x_at	Homo sapiens	compound:'PD 0332991 1 micromolar' vs 'vehicle'	🟦
BRCA1	A_23_P207400	Homo sapiens	growth condition: 'Apical anaerobic' vs. 'Conventional' on A-AGIL-28	🟦
BRCA1	204531_s_at	Homo sapiens	genotype:'HRas over expression' vs 'empty vector control' on A-AFFY-44	🟥

A mock-up of baseline Expression Atlas experiment page, including protein expression data from two external resources, e.g. PRIDE and Peptide Atlas



Showing 50 of 586 genes found:

Gene	adipose	adrenal	brain	breast	colon	heart	kidney	leukocyte	liver	lung	lymph node	ovary	prostate	skeletal muscle	testis	thyroid
⊞ MYL7						581										
⊟ TNNI3						1779					1			2		1
Q01449 PRIDE Q						7										
Q01449 PeptideAtlas														1		
⊞ MYBPC3						347										
⊞ NPPA						338										
⊞ TNNT2	1	0.9	15			3288	22	0.7		7		0.8	14	3	2	1

Expression Level Breakdown for MYL7 (10 transcripts) in heart ?

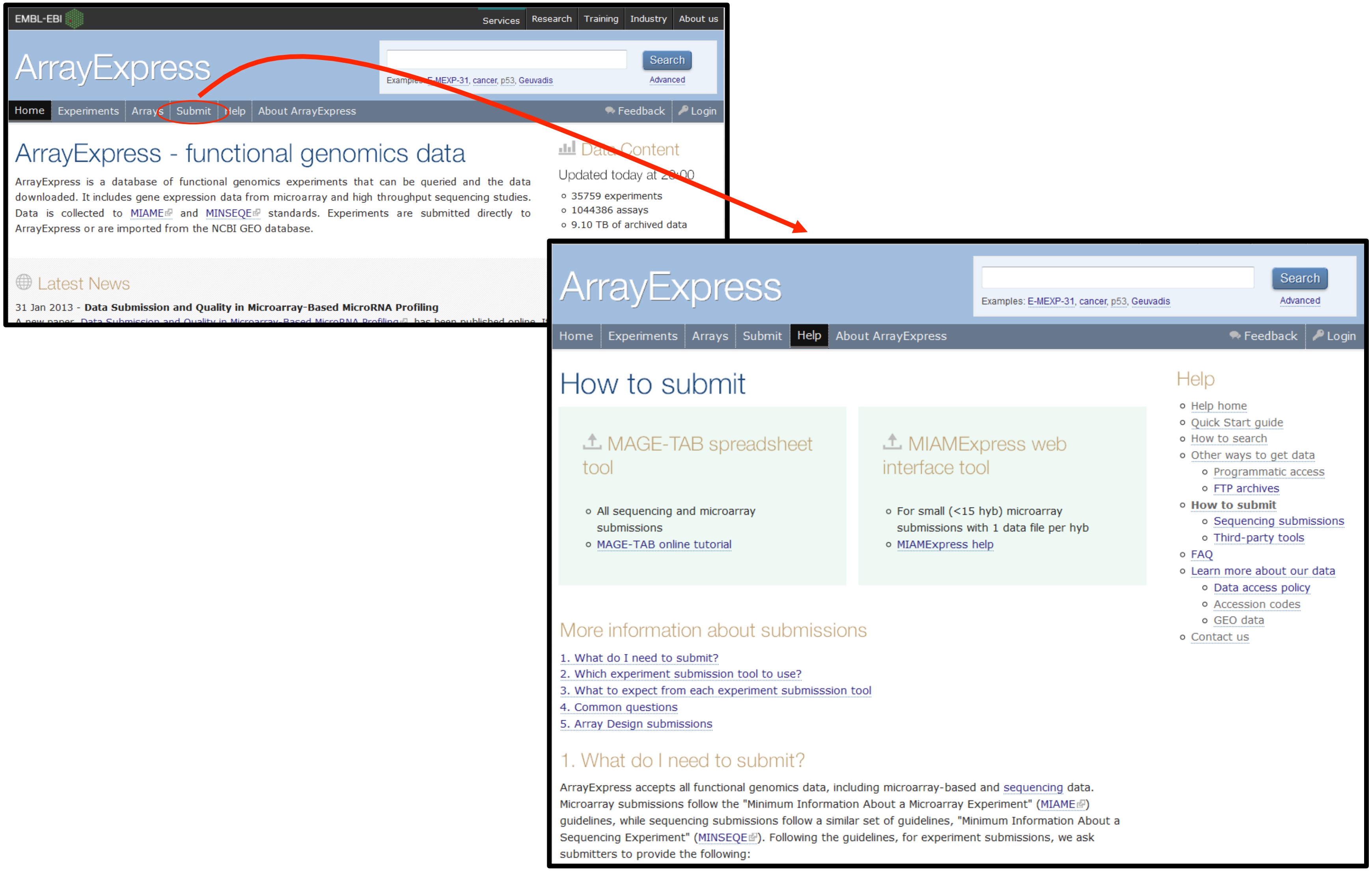
- 🟨 ENST00000223364
- 🟦 ENST00000457314

Number of proteomics experiments in PRIDEQ reporting expression of Q01449 in heart

Dominant transcript (Q01449)

Data submission to ArrayExpress Archive

Data submission to Arrayexpress



ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Data Content

Updated today at 20:00

- 35759 experiments
- 1044386 assays
- 9.10 TB of archived data

Latest News

31 Jan 2013 - **Data Submission and Quality in Microarray-Based MicroRNA Profiling**

How to submit

MAGE-TAB spreadsheet tool

- All sequencing and microarray submissions
- MAGE-TAB online tutorial

MIAMExpress web interface tool

- For small (<15 hyb) microarray submissions with 1 data file per hyb
- MIAMExpress help

More information about submissions

1. What do I need to submit?
2. Which experiment submission tool to use?
3. What to expect from each experiment submisssion tool
4. Common questions
5. Array Design submissions

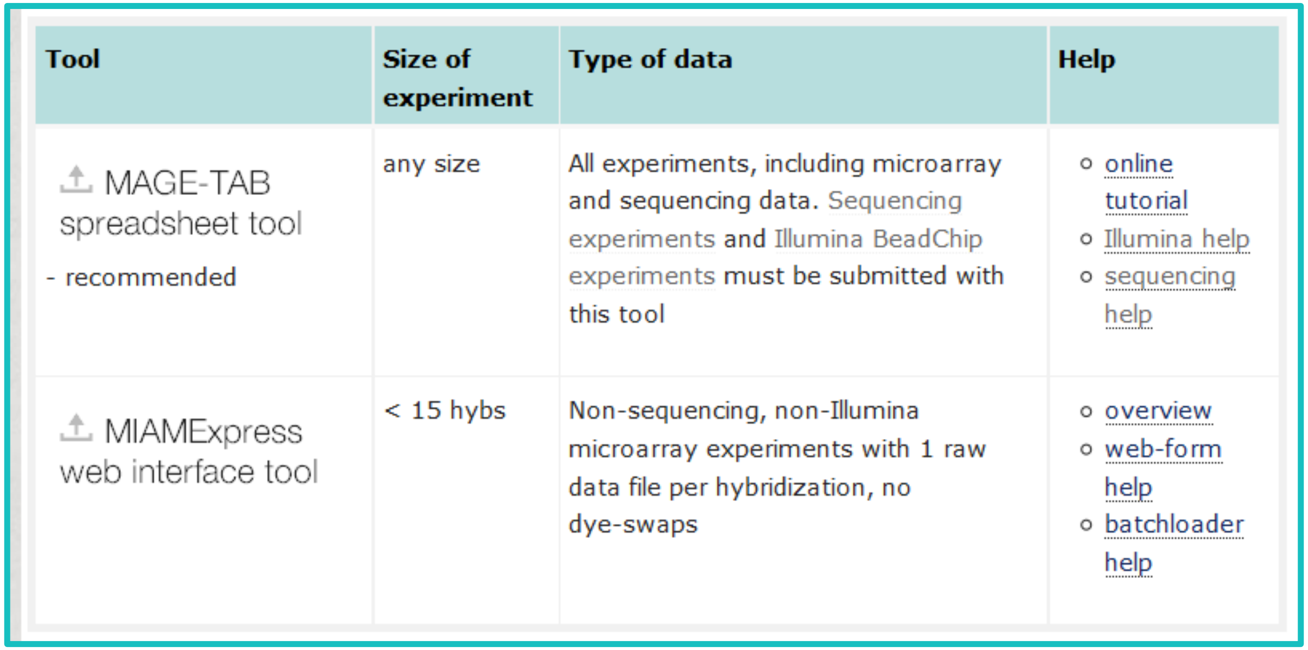
1. What do I need to submit?

ArrayExpress accepts all functional genomics data, including microarray-based and sequencing data. Microarray submissions follow the "Minimum Information About a Microarray Experiment" (MIAME) guidelines, while sequencing submissions follow a similar set of guidelines, "Minimum Information About a Sequencing Experiment" (MINSEQE). Following the guidelines, for experiment submissions, we ask submitters to provide the following:

Help

- Help home
- Quick Start guide
- How to search
- Other ways to get data
  - Programmatic access
  - FTP archives
- **How to submit**
  - Sequencing submissions
  - Third-party tools
- FAQ
- Learn more about our data
  - Data access policy
  - Accession codes
  - GEO data
- Contact us

Data submission to ArrayExpress



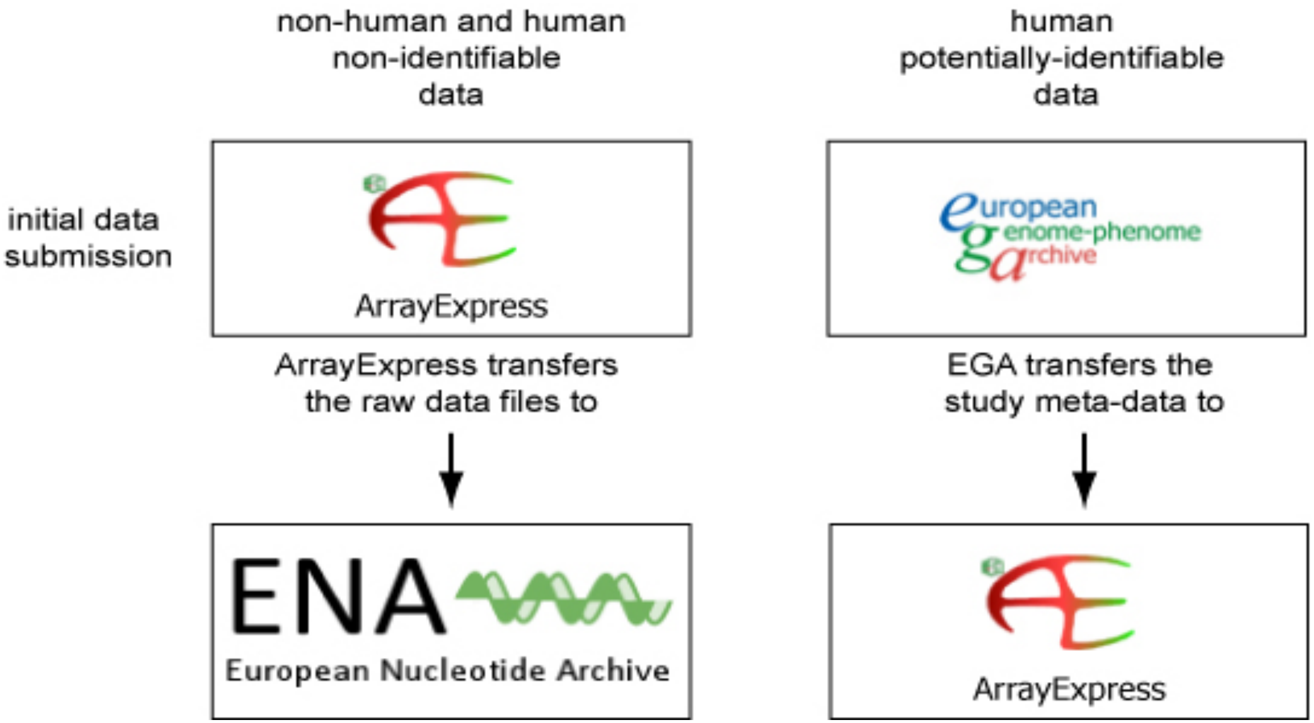
Tool	Size of experiment	Type of data	Help
MAGE-TAB spreadsheet tool - recommended	any size	All experiments, including microarray and sequencing data. Sequencing experiments and Illumina BeadChip experiments must be submitted with this tool	○ online tutorial ○ Illumina help ○ sequencing help
MIAMExpress web interface tool	< 15 hybs	Non-sequencing, non-Illumina microarray experiments with 1 raw data file per hybridization, no dye-swaps	○ overview ○ web-form help ○ batchloader help

- MAGE-TAB route recommended for large/complicated experiments.
- MAGE-TAB template spreadsheet (IDF and SDRF) ***tailor-made*** for your experiment if you follow the MAGE-TAB submission tool

Submission of HTS data

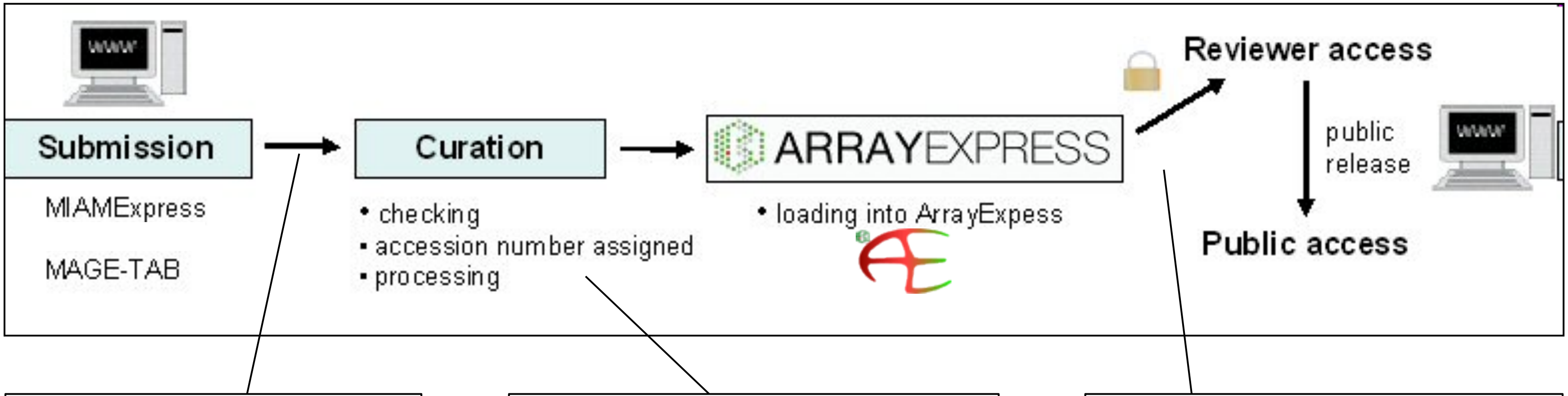
ArrayExpress acts as a “broker” for submitter.

- Meta-data and processed data: ArrayExpress
- Raw sequence reads* (e.g. fastq, bam): ENA



*See http://www.ebi.ac.uk/ena/about/sra_data_format for accepted read file format

What happens after submission?



Submission

MIAMExpress

MAGE-TAB

Curation

- checking
- accession number assigned
- processing

ARRAYEXPRESS

- loading into ArrayExpess

Reviewer access

public release

Public access

Email confirmation

Submission ‘closed’ so no more editing on your end

Curation:

We will email you with any questions

May ‘re-open’ submission for you to make changes

Can keep data private until publication.

Will provide login account details to you and reviewer for private data access

Get your submission in the best possible shape to shorten curation and processing time!

Find out more

- Visit our eLearning portal, **Train online**, at http://www.ebi.ac.uk/training/online/ for courses on ArrayExpress and Atlas
- Watch this short YouTube video on how to navigate the MAGE-TAB submission tool: http://youtu.be/KVpCVGpjw2Y
- Email us at: miamexpress@ebi.ac.uk
- Atlas mailing list: arrayexpress-atlas@ebi.ac.uk