Functional genomics resources at the EBI

Gabriella Rustici
Functional Genomics Group

gabry@ebi.ac.uk
What is functional genomics?

• The aim of functional genomics is to understand the function of genes and other parts of the genome.

• Functional genomic experiments typically utilize genome-wide assays to measure and track many genes (or proteins) in parallel under different environmental 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?
Functional genomics resources at the EBI

- Direct submissions
- Import from external databases (GEO)

ArrayExpress Archive

Curation
Statistical analysis

Expression Atlas

Links to other EBI databases and analysis software, i.e. Bioconductor
What is the difference?

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

Expression Atlas
- Central object: gene/biological condition
- Query for gene expression changes across experiments and across platforms
ArrayExpress Archive – when to use it?

• Find functional genomics 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.
ArrayExpress Archive – what we provide?

• Is a public repository for functional genomics data, mostly generated using microarray or HTS assays

• Serves the scientific community as an archive for data supporting publications, together with GEO at NCBI and CIBEX at DDBJ

• Provides easy access to well annotated data in a structured and standardized format; based on community standards

• Facilitates the sharing of experimental information (e.g. microarray designs, experimental protocols,…)

ArrayExpress Archive content

- HTS data in ArrayExpress and Atlas

![Bar chart showing ArrayExpress Archive content](chart.png)
ArrayExpress scenario

• Imagine that you have just started a project working on **prostate adenocarcinoma**

• You want to find out all the experiments, in the ArrayExpress Archive, that use **RNA sequencing assays** to study this cancer subtype.

• Go to [http://www.ebi.ac.uk/arrayexpress/](http://www.ebi.ac.uk/arrayexpress/)
Browsing the ArrayExpress Archive
Browsing the AE Archive

- Use “prostate adenocarcinoma” to search experiments
- Use filters to narrow down the results
## Browsing the AE Archive

Experiment, citation, sample and factor annotations [clear]
"prostate adenocarcinoma"

Filter on [reset]
All species

Display options [reset]
25 experiments per page

ArrayExpress data only
Advanced query syntax

Submitter/reviewer login
ArrayExpress Browser Help

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Assays</th>
<th>Species</th>
<th>Date</th>
<th>Processed</th>
<th>Raw</th>
<th>Atlas</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOD-24284</td>
<td>Deep sequencing analysis of transcription-induce</td>
<td>14</td>
<td>Homo sapiens</td>
<td>2011-01-19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-GEOD-24283</td>
<td>Deep transcriptional sequencing analysis of human</td>
<td>8</td>
<td>Homo sapiens</td>
<td>2011-01-18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 experiments, 22 assays.
**ArrayExpress Archive – experimental details**

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Assays</th>
<th>Species</th>
<th>Date</th>
<th>Processed</th>
<th>Raw</th>
<th>Atlas</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEO2-24284</td>
<td>Deep sequencing analysis of transcription-induced chimeras in human prostate adenocarcinoma</td>
<td>14</td>
<td>Homo sapiens</td>
<td>2011-01-19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Description**

Prostate adenocarcinoma and matched adjacent normal samples were profiled by deep transcriptional sequencing to analyze transcription-induced chimeras and gene fusions. Reference samples from the MAQC and brain and universal reference libraries were also sequenced. Two-condition experiment plus reference samples: Prostate adenocarcinoma versus matched normal from three separate patients, plus brain and universal reference samples from the MAQC project.

**MINSEQE score**

<table>
<thead>
<tr>
<th>Expt. design</th>
<th>Protocols</th>
<th>Factors</th>
<th>Processed data</th>
<th>Sequence Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

**Contacts**

Thomas D. Wu <twu@gene.com>, Gary Schroth, Irina Khrebtukhova, Somasekar Seshagiri, Thomas D Wu

**Links**

Query Gene Expression Atlas
GEO - GSE24283, ENA - SRP003611
Experimental protocols

**Files**

Investigation Description
Sample and Data Relationship
Browse all available files

**Experiment type**

transcription profiling by high throughput sequencing

**Experimental factors**

<table>
<thead>
<tr>
<th>Factor name</th>
<th>Factor values</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORGAN</td>
<td>brain, Multiple, prostate</td>
</tr>
<tr>
<td>SAMPLEID</td>
<td>Ambion 6050, HF-23260-(1), HF-23261-(1), HF-23732, HF-23751-(1), HF-24126-(1), HF-24127-(1), Stratagene 740000</td>
</tr>
<tr>
<td>SUBTYPE</td>
<td>Adjacent normal prostate, Brain reference, Prostate adenocarcinoma, Universal reference</td>
</tr>
</tbody>
</table>

**Sample attributes**

<table>
<thead>
<tr>
<th>Attribute name</th>
<th>Attribute values</th>
</tr>
</thead>
<tbody>
<tr>
<td>organ</td>
<td>brain, Multiple, prostate</td>
</tr>
<tr>
<td>Organism</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>sampleID</td>
<td>Ambion 6050, HF-23260-(1), HF-23261-(1), HF-23732, HF-23751-(1), HF-24126-(1), HF-24127-(1), Stratagene 740000</td>
</tr>
<tr>
<td>subtype</td>
<td>Adjacent normal prostate, Brain reference, Prostate adenocarcinoma, Universal reference</td>
</tr>
</tbody>
</table>

**Last updated on**

2011-09-14

**Released on**

2011-01-18
ArrayExpress – two databases

Direct submissions
Import from external databases (GEO)

ArrayExpress Archive

Curation
Statistical analysis

Expression Atlas

Links to other EBI databases and analysis software, i.e. Bioconductor

HTS data in ArrayExpress and Atlas
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. Gene Ontology 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 – what we provide?

• Information on gene expression patterns within different biological conditions

• Simple interface to search for gene expression changes measured in various cell types, organism parts, developmental stages, disease states, and many other biological/experimental conditions

• Easy access to meta-analysis results for biologists that don’t have the expertise necessary to analyze vast amount of functional genomics data
Expression Atlas

The ArrayExpress Archive is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. Gene Expression Atlas contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

Experiments Archive
15782 experiments, 442575 assays

Gene Expression Atlas
5670 experiments, 138915 assays, 18346 conditions

News
- 20 Oct 2010 - Internship for a student project in human gene expression - Filled now
  This student project is now taken.
- 17 Sep 2010 - New Atlas Data Release 10.8
  A new release of Gene Expression Atlas has been made with 93 new experiments...more

Links
- ArrayExpress User Survey
- Old ArrayExpress Interface
- Help | Training | FAQ | Citing
- Submit Data (array based and re-sequencing)
- Programmatic Access | FTP Access
- Software Downloads and Statistics
- EFO | Bioconductor Package | Quality Metrics
- ArrayExpress Scientific Advisory Board
- Functional Genomics Group

Gene Expression Atlas Home
Expression Atlas home page

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

Query for genes

Restrict query by direction of differential expression

Query for conditions

The ‘advanced query’ option allows building more complex queries.
Expression Atlas scenario

Imagine you are a scientist working in a drug discovery lab developing new therapies for neurodegenerative diseases.

You want to find human genes involved in the disease that could possibly be targets for drug therapy.

You have recently read a paper stating that 'glutamate receptors are important in neurodegeneration'.

So you are particularly interested in finding signaling receptor proteins containing an NMDA domain (a particular class of glutamate receptor) that are deregulated in neurodegenerative disease.
Search Atlas for genes

- Interested in genes involved in ‘receptor activity’
- Use GO term ‘receptor activity’ in genes search box
Add experimental conditions to the search

- search for ‘nervous system disease’
Search results – heatmap view

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>ACVR1B</th>
<th>KDELR2</th>
<th>CD44</th>
<th>JMJD6</th>
<th>SLC7A1</th>
<th>CD47</th>
<th>NR2F2</th>
<th>NTRK3</th>
<th>YKT6</th>
<th>IL6ST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9</td>
<td>4</td>
<td>10</td>
<td>11</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Parkinson's disease</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Brain tumors</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Central nervous system cancer</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Granulocytosis</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Hereditary spastic paraplegia</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hypertension</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Metabolic disorder</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Mental or Behavioural disorder</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Multiple sclerosis, progressive muscular atrophy, spinal muscular atrophy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Multiple sclerosis, progressive spastic paraplegia</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Spinal muscular atrophy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tumour of cranial bone and spine</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Advance query

- search if among these genes there are some which encode for a protein carrying an NMDA receptor domain
GRIN1 is differentially expressed in 222 experiments: 136 organism parts: prostate, amygdala, ... (134 more), 203 disease states: normal, prostate carcinoma, ... (201 more), 76 cell types 293 cell lines 29 compound treatments and 86 other conditions.

Synonyms
NMDAR1

Orthologs
ENSCRGO0000009574 (Cliona intestinalis) Grin1 (Rattus norvegicus) Q6R612 (Gallus gallus) Grin1 (Mus musculus) grin1b (Danio rerio) grin1a (Danio rerio) AGAP001478 (Anopheles gambiae) nmr1 (Caenorhabditis elegans) Nmdar1 (Drosophila melanogaster) (Compare orthologs)

InterPro Term
GPCR_3, S5P_bac_3, trans.glu_rct, NMDA_rct, ANF_tap-bd_rct

Gene Ontology Term
transporter activity, plasma membrane, ionotropic glutamate receptor signaling pathway, transport, long-term memory, protein binding, regulation of excitatory postsynaptic membrane potential, regulation of axonogenesis, calcium channel activity, male mating behavior, visual learning, cytoplasm, memory, membrane, postsynaptic membrane, ion channel activity, social behavior, neuromuscular process, learning or memory, regulation of memubranaton, cation transport, receptor binding, ionotropic glutamate receptor activity, membrane fraction, synaposome, sensory perception of pain, cell junction, regulation of synaptic plasticity, glutamate binding, positive regulation of apoptosis, adult locomotory behavior, synaptic vesicle, learning, N-methyl-D-aspartate selective glutamate receptor complex, calmodulin binding, positive regulation of transcription from RNA polymerase II promoter, cation channel activity, cellular calcium ion homeostasis, integral to plasma membrane, conditioned taste aversion, extracellular-glutamate-gated ion channel activity, postsynaptic density, dendrite, dendritic spine, receptor activity, outer membrane-bounded periplasmic space, synapse, calcium ion transport, olfactory learning, regulation of dendrite morphogenesis, response to ethanol, startle response, synaptic transmission, glutamatergic, calcium ion binding, associative learning, calcium ion homeostasis, negative regulation of neuron apoptosis, regulation of long-term neuronal synaptic plasticity, regulation of neuronal synaptic plasticity, regulation of respiratory gaseous exchange, cerebral cortex development, regulation of neuron apoptosis, regulation of cell communication, response to morphine, respiratory gaseous exchange, response to anaphetamine, N-methyl-D-aspartate selective glutamate receptor activity, suckling behavior, pons maturation, prepulse inhibition, regulation of synaptogenesis, glycine binding

UniProt Accession
Q55596, Q9UPF8, Q9UPF0, A2AVK2, Q5VSF0, Q5VSF5, Q5VSF4, Q59H41, Q59GW0

Search EB-eye
ENSG00000176084

Show more properties

---

**Experimental Factors**

Organism part
- Studied in E-GEOD-24283, E-MTAB-26, E-APMY-5, E-TABM-302, E-GEOD-802, ... (23 experiments)
- Number of published studies where the gene is over/under expressed compared to the gene's overall mean expression level in the study.

Cell line
- Studied in E-GEOD-17482, E-GEOD-5258, E-MEXP-1006, E-GEOD-19639, E-GEOD-18005, ... (29 experiments)

**Expression Profiles**

222 experiments showing differential expression

E-MTAB-62: Human gene expression atlas of 5372 samples representing 369 different cell lines

<table>
<thead>
<tr>
<th>Factor Value</th>
<th>u/o</th>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SKBR3</td>
<td>2</td>
<td>E-MTAB-62, E-MEXP-440</td>
</tr>
<tr>
<td>MOLT4</td>
<td>2</td>
<td>E-MTAB-62, E-MEXP-440</td>
</tr>
<tr>
<td>SW620</td>
<td>2</td>
<td>E-MEXP-1014, E-MEXP-37</td>
</tr>
<tr>
<td>primary culture</td>
<td>2</td>
<td>E-GEOD-1657, E-GEOD-2450</td>
</tr>
<tr>
<td>MDA-MB-231</td>
<td>2</td>
<td>E-TABM-423, E-GEOD-5823, E-MTAB-62</td>
</tr>
<tr>
<td>BT474</td>
<td>2</td>
<td>E-GEOD-1657, E-GEOD-440</td>
</tr>
</tbody>
</table>

267 more value(s).

Show expression data for all values of this factor >>

Array Design: A-AFFY-33 Affymetrix GeneChip
- Human Genome HG-U133A (HG-U133A)
- Show expression profile / experiment details

E-MTAB-25: Transcription profiling of human and mouse organism parts where organism p parts between the two species

Experimental Factors

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

---

Gene page
# Search results – heatmap view

<table>
<thead>
<tr>
<th>Gene</th>
<th>GRIN1</th>
<th>GRIA3</th>
<th>GRIK2</th>
<th>GRIN2A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

*GRIA3 in Huntington's disease* (EFO) overexpressed in 1, underexpressed in 3 and non-differentially expressed in 1 experiment(s).

---

**E-MTAB-62**: Human gene expression atlas of 5372 samples representing 369 different cell and tissue types, disease states and cell lines.

**Disease state**

- Brain tumor: \( \uparrow < 10^{-10} \)
- Huntington's disease: \( \uparrow < 10^{-10} \)
- Bipolar disorder: \( \uparrow < 10^{-10} \)
- Alzheimer's disease: \( \uparrow < 10^{-10} \)
- Acute lymphoblastic leukemia: \( \uparrow 1.61 \times 10^{-9} \)
- Multiple myeloma: \( \uparrow < 10^{-10} \)

*Array Design: A-AFFY-33 Affymetrix GeneChip Human Genome HG-U133A [HG-U133A]*
Experiment view

Transcription profiling of human blood from Huntington's disease patients

(PubMed 16043692)

Data shown for array design: A-AFFY-33

E-GEOD-1751

Platform: A-AFFY-33
Organism: homo sapiens
Samples: 31
ArrayExpress: E-GEOD-1751

- experiment design

<table>
<thead>
<tr>
<th>Gene</th>
<th>Design Element</th>
<th>Experimental Factor</th>
<th>Factor Value</th>
<th>UP/DOWN</th>
<th>T-Statistic</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000125675</td>
<td></td>
<td>All factor values</td>
<td></td>
<td>All expressions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRIA3</td>
<td>206730_at</td>
<td>Disease state</td>
<td>Huntington's disease</td>
<td>DOWN</td>
<td>-3.7</td>
<td>0.013</td>
</tr>
<tr>
<td>GRIA3</td>
<td>208032_s_at</td>
<td>Disease staging</td>
<td>symptomatic</td>
<td>NONDE</td>
<td>-2.5</td>
<td>0.129</td>
</tr>
</tbody>
</table>
Dealing with HTS data

- HTS is becoming an extremely popular technology in functional genomics
- It provides higher power, sensitivity and reproducibility than expression microarrays
- Offers more than microarrays, enabling analysis of alternative transcription start site usage, alternative splicing, alternative polyadenylation and allele-specific expression
- But the ability to analyze and interpret the data can be the limiting step, especially for biologists
HTS data in ArrayExpress Archive

HTS vs other technologies

- HTS experiments: 97%
- Other: 3%

RNA-seq vs DNA-seq

- RNA seq: 7%
- DNA seq: 55%
- RNA seq, DNA-seq: 38%

RNA-seq: coding vs non coding

- Coding: 32%
- Non coding: 68%
Dealing with HTS data

• We have developed an R based pipeline, named ArrayExpressHTS, for pre-processing, expression estimation and data quality assessment of HTS datasets

• Currently handling RNA-seq and ChIP-seq data

• The pipeline can be used for analyzing:
  ✓ private data
  ✓ public data, available through ArrayExpress and ENA

• It can be used:
  ✓ on a local computer
  ✓ remotely on the EBI R Cloud, www.ebi.ac.uk/tools/rcloud
RNA-seq processing pipeline

Direct data submissions and GEO import

ArrayExpress Archive

Expression Atlas

EGA

Expression Atlas

ENA

Ensembl

RNAseq Processing pipeline

SDRF

FastQ

BAMs

RPKM

FASQ files

Short reads (FASTQ files)

Summary level data

Direct data submissions and GEO import

RNA-seq Processing pipeline
ArrayExpressHTS on the R cloud

Pipeline tools:
- tophat
- bowtie
- bwa
- cufflinks
- samtools

ArrayExpressHTS R package

R-server

References, Index & Annotation

User Project Storage

ENAX

- RAW DATA
- Experiment meta data

- ExpressionSet
- Quality reports

- SDRF
- IDF

ArrayExpress

06/10/2011

HTS data in ArrayExpress and Atlas
To find out more

- eLearning courses on Archive and Atlas: http://www.ebi.ac.uk/training/online/


- R-cloud: http://www.ebi.ac.uk/Tools/rcloud/

- Submissions: http://www.ebi.ac.uk/fg/submissions_overview.html