Automating the data import from the ArrayExpress database into Bioconductor

Audrey Kauffmann

ArrayExpress

- Public repository for microarray data supporting Microarray and Gene Expression Database (MGED) standards
- Archival repository for microarray data supporting publications, together with GEO at NCBI (USA) and CIBEX at DDBJ (Japan)
- Provides easy access to well annotated microarray data in a structured and standardized format
- Facilitates the sharing of microarray designs and protocols
- MGED standards: MIAME, MAGE and Ontology

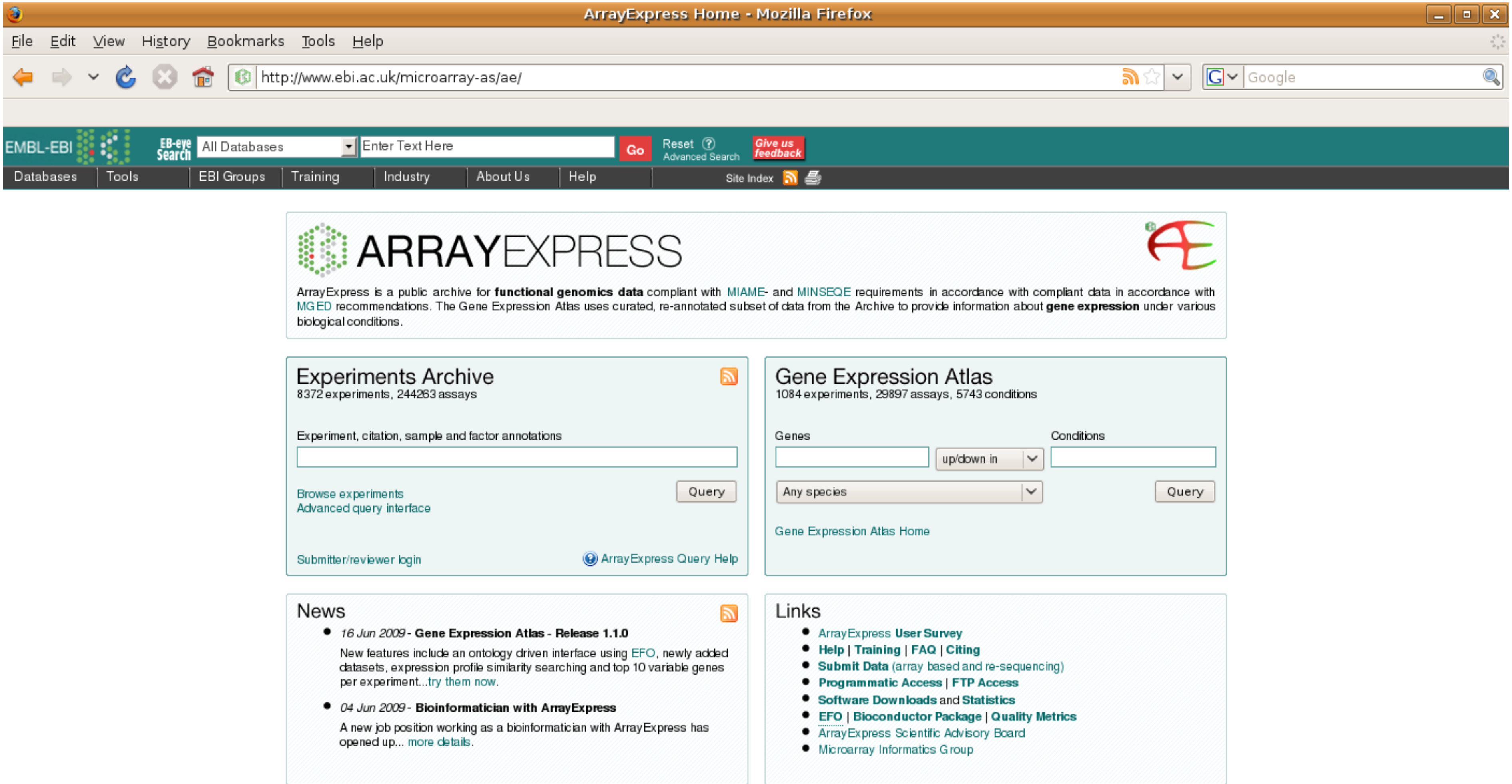
ArrayExpress

www.ebi.ac.uk/arrayexpress

23/07/2009:

8372 experiments

244263 assays



ARRAYEXPRESS

ArrayExpress is a public archive for **functional genomics data** compliant with MIAME- and MINSEQE requirements in accordance with compliant data in accordance with MGED recommendations. The Gene Expression Atlas uses curated, re-annotated subset of data from the Archive to provide information about **gene expression** under various biological conditions.

Experiments Archive

8372 experiments, 244263 assays

Experiment, citation, sample and factor annotations

Query

Browse experiments

Advanced query interface

Submitter/reviewer login

ArrayExpress Query Help

Gene Expression Atlas

1084 experiments, 29897 assays, 5743 conditions

Genes

up/down in

Conditions

Any species

Query

Gene Expression Atlas Home

News

- *16 Jun 2009* - **Gene Expression Atlas - Release 1.1.0**

  New features include an ontology driven interface using EFO, newly added datasets, expression profile similarity searching and top 10 variable genes per experiment...try them now.

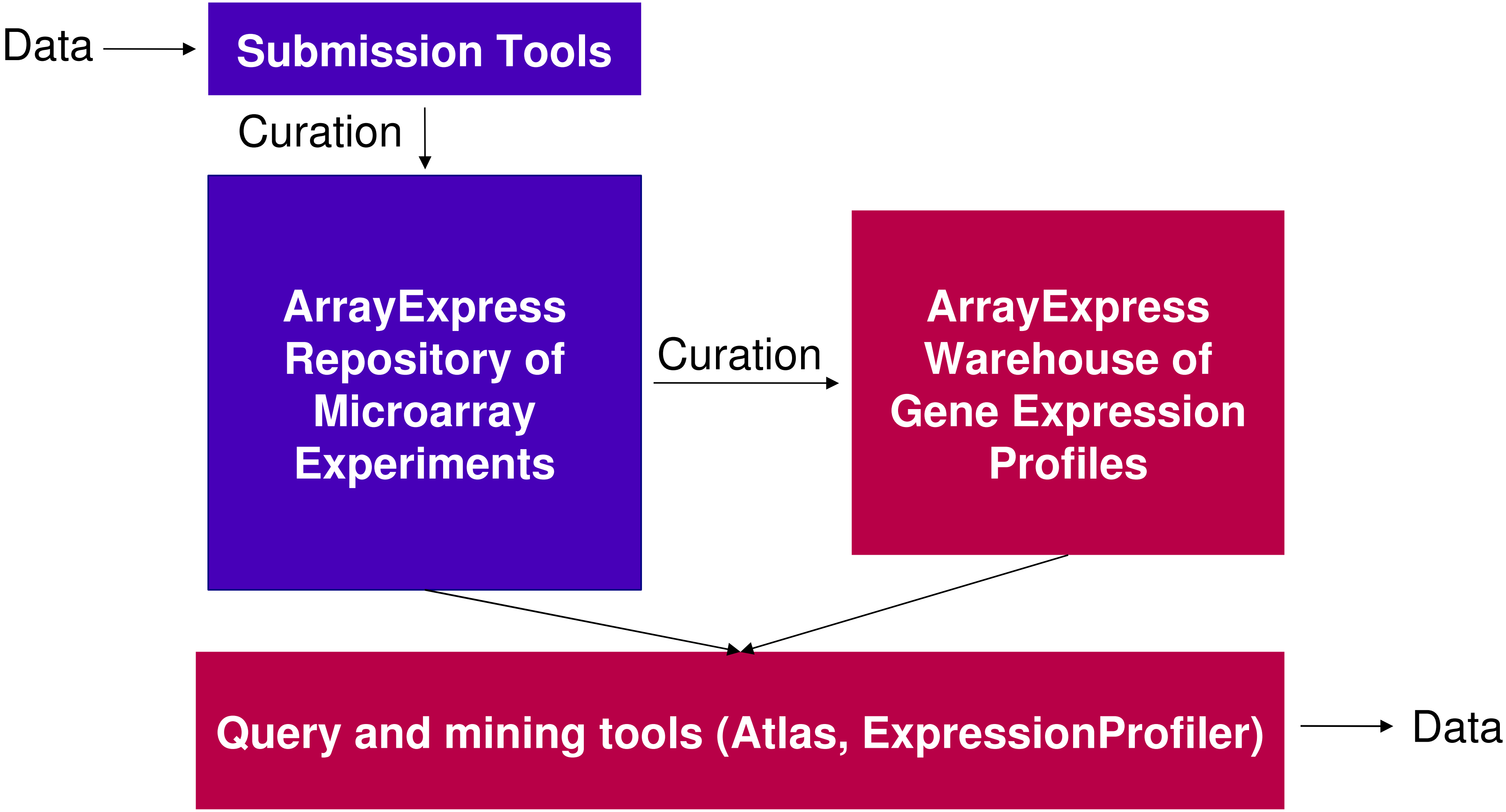
- *04 Jun 2009* - **Bioinformatician with ArrayExpress**

  A new job position working as a bioinformatician with ArrayExpress has opened up... more details.

Links

- ArrayExpress **User Survey**
- **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
- Microarray Informatics Group

ArrayExpress – Two databases



Data

Submission Tools

Curation

ArrayExpress Repository of Microarray Experiments

Curation

ArrayExpress Warehouse of Gene Expression Profiles

Query and mining tools (Atlas, ExpressionProfiler)

Data

MIAME

- www.mged.org/Workgroups/MIAME/miame.html
- Minimal Information About a Microarray Experiment (Brazma et al., 2001 – Nature Genetics)
- Describes what is needed to:
  - Enable the interpretation of the results of the experiment unambiguously
  - Potentially to reproduce the experiment

MIAME compliant data

6 critical elements contributing towards MIAME:

- Essential sample annotation including experimental factors and their values (e.g. compound and dose)
- Experimental design including sample data relationships (e.g. which raw data file relates to which sample)
- Sufficient array annotation (e.g. gene identifiers, genomic coordinates, probe sequences or array catalog number)
- Essential laboratory and data processing protocols (e.g. normalization method used)
- Raw data for each hybridization (e.g. CEL or GPR files)
- Final normalized data for the set of hybridizations in the experiment

MIAME in practice

- MIAME does not specify a particular format but MGED recommends the use of MAGE-TAB, which is based on spreadsheets
- MIAME also does not specify any particular terminology, however MGED recommends the use of MGED Ontology for the description of the key experimental concepts or ontologies developed by the respective communities for describing specific terms (http://obofoundry.org/)

MAGE-TAB

5 files:

- SDRF *Sample and Data Relationship Format:* txt file
- ADF *Array Design Format:* txt file
- IDF *Investigation Description File:* txt file
- Raw archive: zip file (containing CEL, GPR...)
- Processed: txt file

Behind ArrayExpress

- Microarray Informatics Team EBI - Alvis Brazma
- ArrayExpress curation team:
  - Helen Parkinson
  - Anna Farne
  - Ele Holloway
  - Margus Lukk
  - Eleanor Williams
  - email: miamexpress@ebi.ac.uk

Bioconductor data structures

- Objects
  - AffyBatch: Affymetrix arrays
  - ExpressionSet: One colour arrays
  - NChannelSet: Two colours arrays
- Structure
  - assayData: expressions
  - phenoData: sample annotation
  - featureData: probes annotation
  - experimentData: MIAME experiment level information
  - annotation: Bioconductor annotation to use
  - cdf: cdf package associated (AffyBatch only)

AffyBatch / ExpressionSet

Expression values (exprs)

	Sample 1	Sample 2	...	Sample i
Probe 1	$I_{1,1}$	$I_{1,2}$	...	$I_{1,i}$
Probe 2	$I_{2,1}$	$I_{2,2}$	...	$I_{2,i}$
...	...	...	...	...
Probe k	$I_{k,1}$	$I_{k,2}$	...	$I_{k,i}$

Probe annotation (featureData)

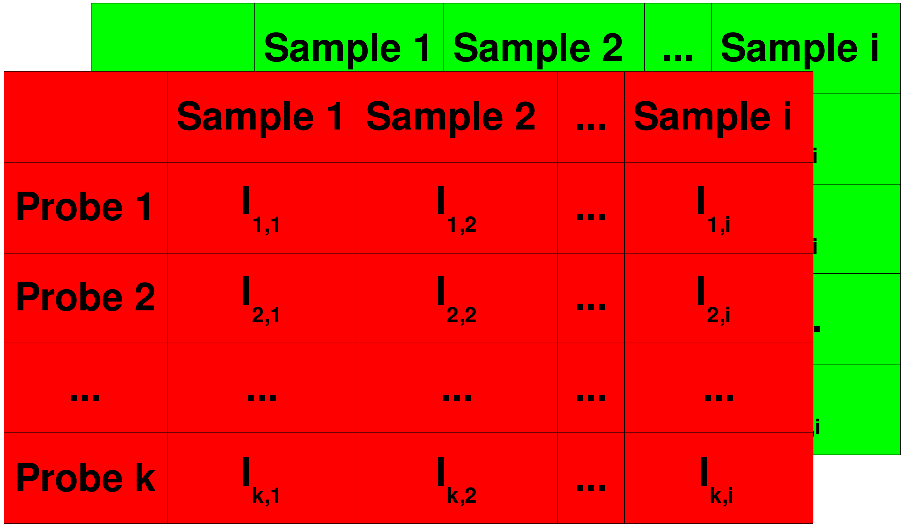
	X	Y	ID	...
Probe 1	1	1	NM_000456	...
Probe 2	2	1	NM_007294	...
...	...	...	...	...
Probe k	244	180	NM_000594	...

Sample annotation (phenoData)

	Cell type	Treatment	Replicate	...
Sample 1	WT	Yes	1	...
Sample 2	WT	Yes	2	...
...	...	...	...	...
Sample i	Mut	No	2	...

NChannelSet

Expression values (assayData)



	Sample 1	Sample 2	...	Sample i
Probe 1	$I_{1,1}$	$I_{1,2}$	...	$I_{1,i}$
Probe 2	$I_{2,1}$	$I_{2,2}$	...	$I_{2,i}$
...	...	...	...	...
Probe k	$I_{k,1}$	$I_{k,2}$	...	$I_{k,i}$

Probe annotation (featureData)

	X	Y	ID	...
Probe 1	1	1	NM_000456	...
Probe 2	2	1	NM_007294	...
...	...	...	...	...
Probe k	244	180	NM_000594	...

Sample annotation (phenoData)

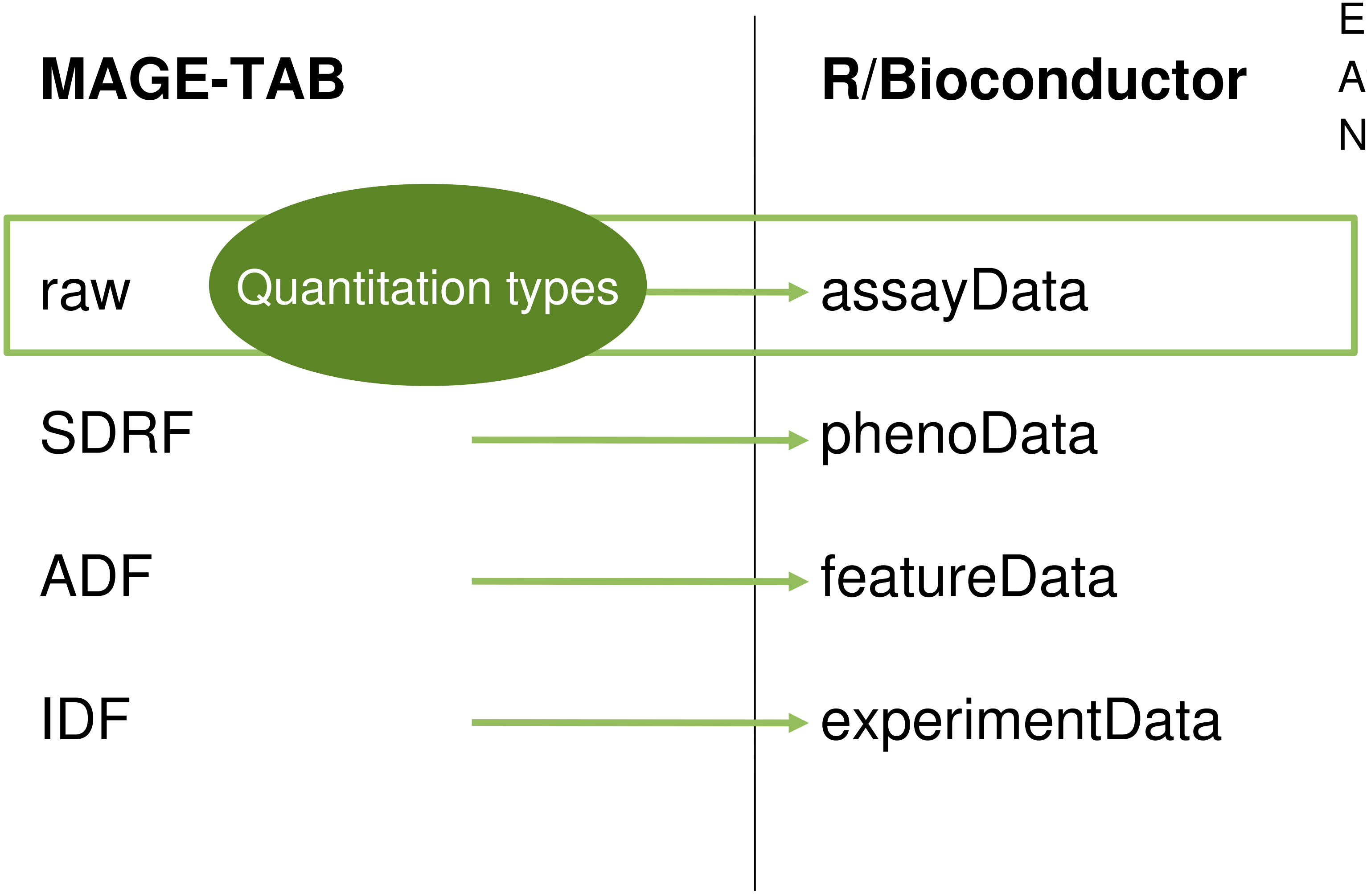
	Cell type	Treatment	Replicate	...
Sample 1	WT	Yes	1	...
Sample 2	WT	Yes	2	...
...	...	...	...	...
Sample i	Mut	No	2	...

ArrayExpress Bioconductor package

Goal:

- Provide easy conversion of ArrayExpress data sets into R objects for further analyses
- Meta-analysis
- Comparison
- Validation

ArrayExpress Bioconductor package



MAGE-TAB

R/Bioconductor

raw

Quantitation types

assayData

SDRF

phenoData

ADF

featureData

IDF

experimentData

ExpressionSet
AffyBatch
NChannelSet

ArrayExpress package functions

- `queryAE`: query the database
- `ArrayExpress`: build object from raw data
- `getAE`: download MAGE-TAB files
- `magetab2bioc`: convert MAGE-TAB files (local or from the database) into an object
- `getcolproc`: extracts the column names from processed MAGE-TAB
- `procset`: converts local MAGE-TAB files into an ExpressionSet

Hands-on