International Efforts Towards Standardizing the Exchange of Microarray Experiment Information and Data

Pierre Bushel
NIEHS Microarray Group Bioinformatics Manager
bushel@niehs.nih.gov

November 15th 2002

NIEHS Microarray Users Group meeting

Presentation Outline

"The what, why and how pertaining to a standard for microarray information and data"

- Microarray Gene Expression experiment and data specifications
- Microarray data exchange ideology
- NIEHS Microarray Group and other information and data exchange methods

“What is the standard for microarray information and data ?”

MIAME

Minimal Information About a Microarray Experiment

A detailed specification of the minimal information needed to be captured and tracked in order to unify the annotation of a microarray experiment

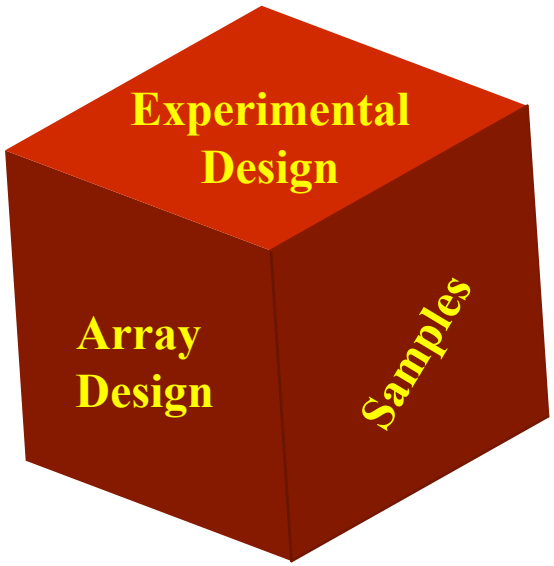
Since 1999 the specification of MIAME has be crafted and refined by a Microarray Gene Expression Database (MGED) working group of international scientists under the leadership of Alvis Brazma, team leader of the Microarray Informatics Group at the European Bioinformatics Institute http://www.mged.org

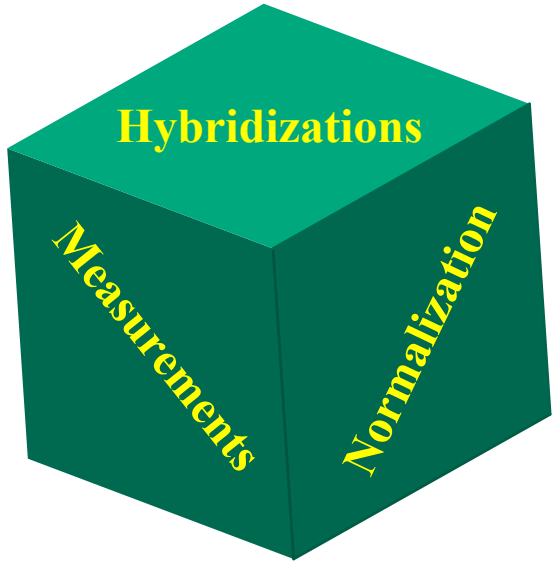
The MIAME specification was widely accepted as a standard by conference attendees at the MGED3 meeting, Stanford University, March 28th, 2001.

MIAME 1.1 Draft 6

http://www.mged.org/Workgroups/MIAME/miame_1.1.html

MIAME Components

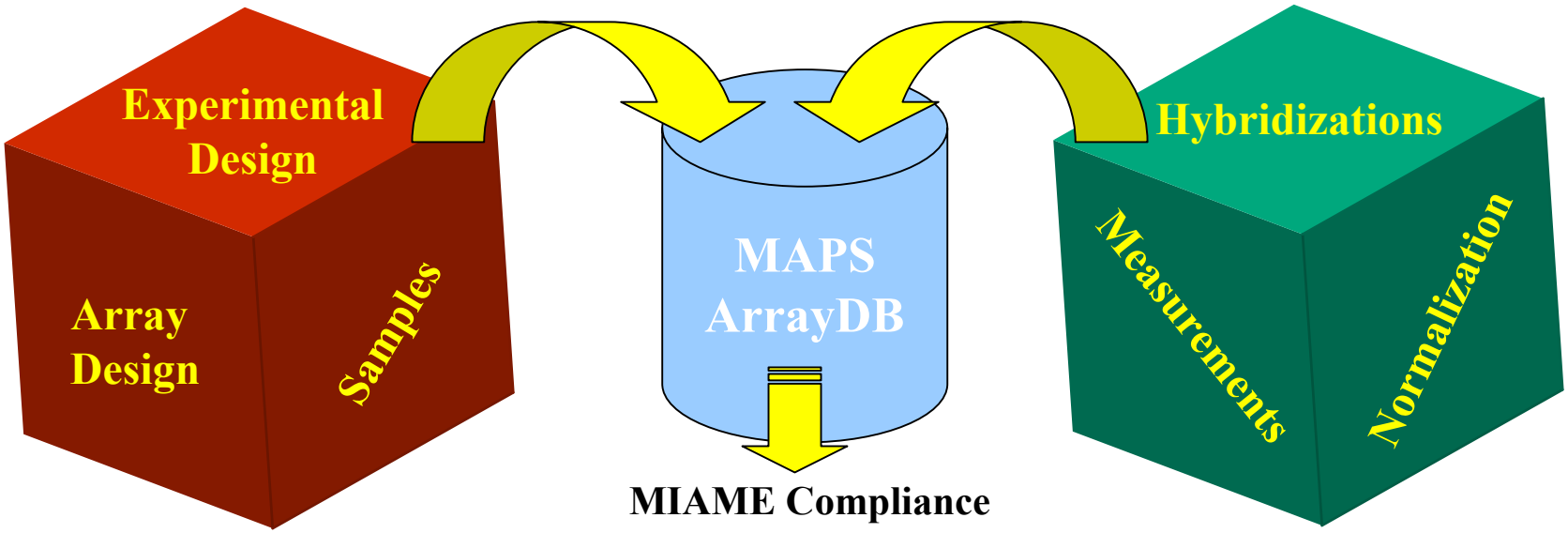




- Submitter information
- Experimental parameters
- Array Platform used
- Experiment series details
- Description of array
- Description of element
- Description of element properties – DNA sequence
- Biological sample source and treatment
- Sample preparation

- Hybridization protocol and parameters
- Washing, and blocking details
- Digitized raw image file
- Scanning specifications
- Image analysis and quantification
- Complete image analysis output
- Replicate information
- Normalization strategy and algorithm
- Control array elements

NIEHS Microarray Group Database Compliance



MIAME Compliance

- ✓Submitter information
- ✓Experimental parameters
- ✓Array Platform used
- ✓Experiment series details
- ✓Description of array
- ✓Description of element
- ✓Description of element properties – DNA sequence
- ✓Biological sample source and treatment
- ✓Sample preparation*

- ✓Hybridization protocol and parameters*
- ✓Washing, and blocking details*
- ✓Digitized raw image file
- ✓Scanning specifications*
- ✓Image analysis and quantification
- ✓Complete image analysis output
- ✓Replicate information
- ✓Normalization strategy and algorithm
- ✓Control array elements

*More detail can be stored in the MAPS/ArrayDB database

“Why the standard for microarray information and data ?”

Latest Breaking News!

4th MGED Programming Jamboree 29/10/2002

The 4th MGED programming jamboree will take place between 06/12/2002 and 10/12/2002 at the Stanford campus. For more details, check here.

First MGED Sponsor 28/10/2002

Iobion becomes the first official sponsor of MGED.

MIAME Checklist Endorsement 14/10/2002

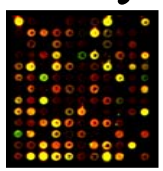
Major scientific journals (including Cell, The Lancet, Nature and Science) endorse MIAME guidelines for publishing microarray gene expression data.

MAGE-ML Paper Published 14/10/2002

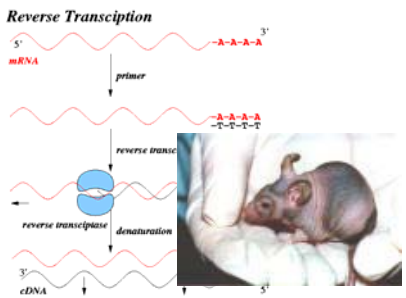
A paper by P.Spellman et al, "Design and implementation of microarray gene expression markup language (MAGE-ML)" has been published in Genome Biology (issue 9), documenting the results of the collaborative effort to establish a standard for gene expression data exchange; see here.

The goal of MIAME is to outline the minimum information required to interpret unambiguously and potentially reproduce and verify an microarray based gene expression monitoring experiment

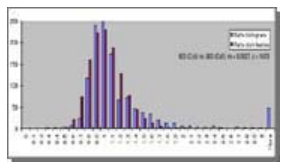
Array Platforms



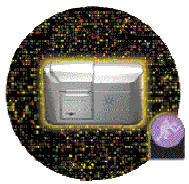
Protocols



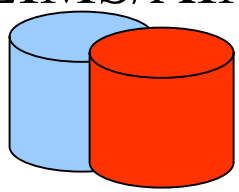
Data Acquisition Software



Scanning Technology

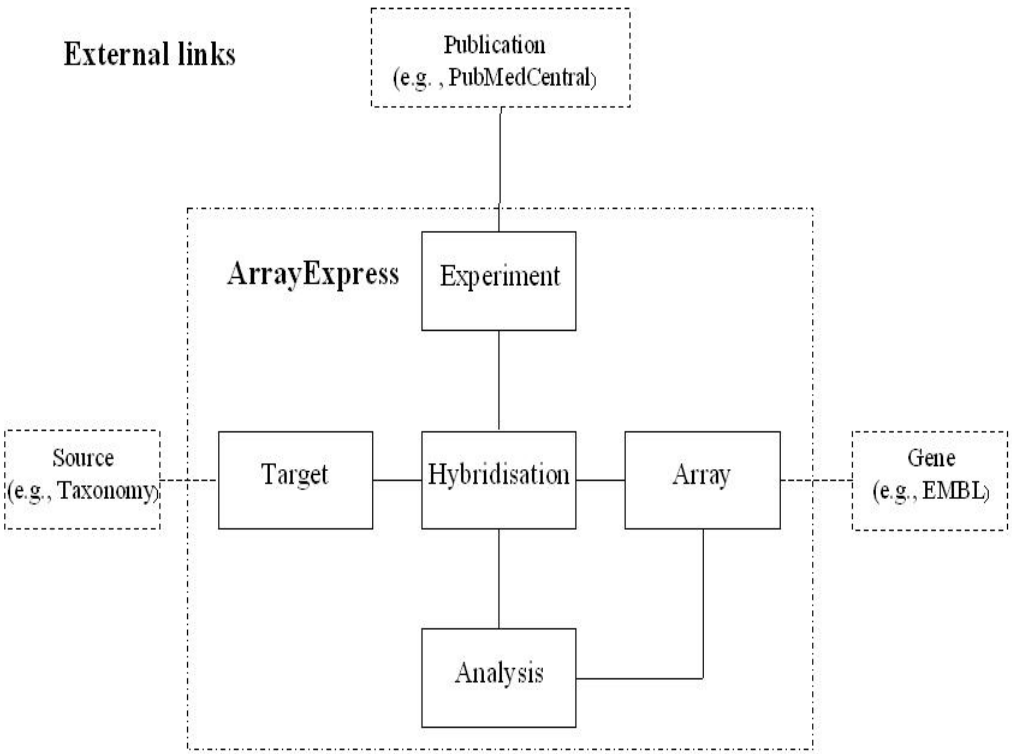


LIMS/AIMS



Public Gene Expression Databases

Top level structure of ArrayExpress database



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

The GeneX project consists of 4 parts:

Curation Tool (Java-based) - to get the user's data into the database.

Database - currently GeneX is being run on the Open Source PostgreSQL database system.

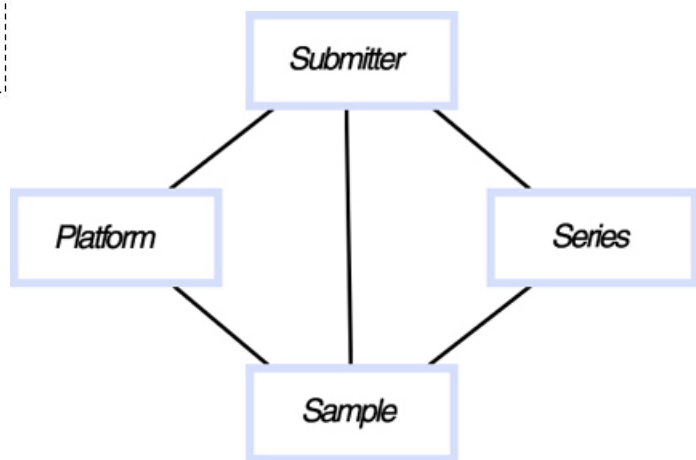
XML Data Exchange Protocol - (NCGR's is called GeneXML

Query and Analytical Routines - to search for and analyze data from the database

http://www.ncgr.org/genex/

Gene Expression Omnibus

Database four entities



Warehouses filter, oligoarrays, microarray, or SAGE data from dose-response, time course, repeated, and parallel platform analyses

http://www.ncbi.nlm.nih.gov/geo/

“How to comply or support the standard for microarray information and data ?”

MAGE-ML

MicroArray Gene Expression – Markup Language

An Extensible Markup Language with tags to wrap microarray gene expression information and data for improved human comprehension and interpretation in web environments and/or database system

Since 1999 the specification of MAGE-ML has be crafted and refined by a Microarray Gene Expression Database (MGED) working group of international scientists under the leadership of Paul Spellman, Scientist at the University of California at Berkeley

The MAGE-OM was accepted as an object by the Object Management Group in February, 2001

OMG is an open membership, not-for-profit consortium that produces and maintains computer industry specifications for interoperable enterprise applications.

XML Technology Syntax

- **Tag: String of text determining the opening or closing component of an element**
- **Attribute: Name and value in an element's opening tag**
- **Element: The complete object defined by open and closed tags, attributes, and sub-elements**

```
<?xml version =“1.0”?>

<investigator>
        <project>
                <experiment>Rat 99-01864</experiment>
                <agent>Wyeth</agent>
         </project>

        <name>
                <firstname>Hisham</firstname>
                <lastname>Hamadeh</lastname>
        </name>
        <address type =“work”>
                <street>NIEHS 111 Alexander Drive</street>
                <city>RTP</city>
                <state>NC</state>
        </address>
        <comments>Interest in Rat Peroxisome Proliferators</comments>
</investigator>
```

MAGE-ML Quantitation_Type Package


```
  <!DOCTYPE MAGE-ML (View Source for full doctype...)>
- <MAGE-ML identifier="Protocol_test">
  - <QuantitationType_package>
    - <QuantitationType_assnlist>
      - <MeasuredSignal identifier="Intensity" name="Affymetrix Intensity" isBackground="false">
        - <Scale_assn>
            <OntologyEntry category="MGED:Scale" value="Linear" />
          </Scale_assn>
        - <DataType_assn>
            <OntologyEntry category="MGED:DataType" value="float" />
          </DataType_assn>
        </MeasuredSignal>
      - <MeasuredSignal identifier="CH1I" name="Cy3 Foreground" isBackground="false">
        - <Channel_assnref>
            <Channel_ref identifier="CH1" />
          </Channel_assnref>
        - <Scale_assn>
            <OntologyEntry category="MGED:Scale" value="Linear" />
          </Scale_assn>
        - <DataType_assn>
            <OntologyEntry category="MGED:DataType" value="float" />
          </DataType_assn>
        </MeasuredSignal>
      - <MeasuredSignal identifier="CH1B" name="Cy3 Background" isBackground="true">
        - <Channel_assnref>
            <Channel_ref identifier="CH1" />
          </Channel_assnref>
        - <Scale_assn>
            <OntologyEntry category="MGED:Scale" value="Linear" />
          </Scale_assn>
        - <DataType_assn>
            <OntologyEntry category="MGED:DataType" value="float" />
          </DataType_assn>
        </MeasuredSignal>
```

MAGE-ML Packages

- Audit and Security – Controls users permissions to view data and annotation
- Description – Allows reference to third party annotation
- Measurement – Provides information on quantities (with a unit)
- BQS – Permits reference to an article, book or publication
- BioEvent – Allows representation of an event to produce a target
- Protocol – Describes lab procedures or analysis algorithms
- BioMaterial – Represents biological material and substances (cells, tissue, DNA, protein)
- BioSequence – Describes genes or sequences

MAGE-ML Packages, continued

- DesignElement – Tracks what is intended to be at each location on the array
- ArrayDesign – Describes the microarray printed and hybridized
- ArrayPackage – Details processes of creating arrays
- BioAssay – Represents physical and computational groupings of arrays and biomaterial
- QuantitationType – Defines quantitation for derivation of signal, error, statistics
- BioAssayData – Contains information and annotation of joining an array with BioMaterial, image acquisition, and data extraction
- Experiment – Hierarchical grouping of BioAssays
- HigherLevelAnalysis – Describes results of higher order analysis on BioAssayData

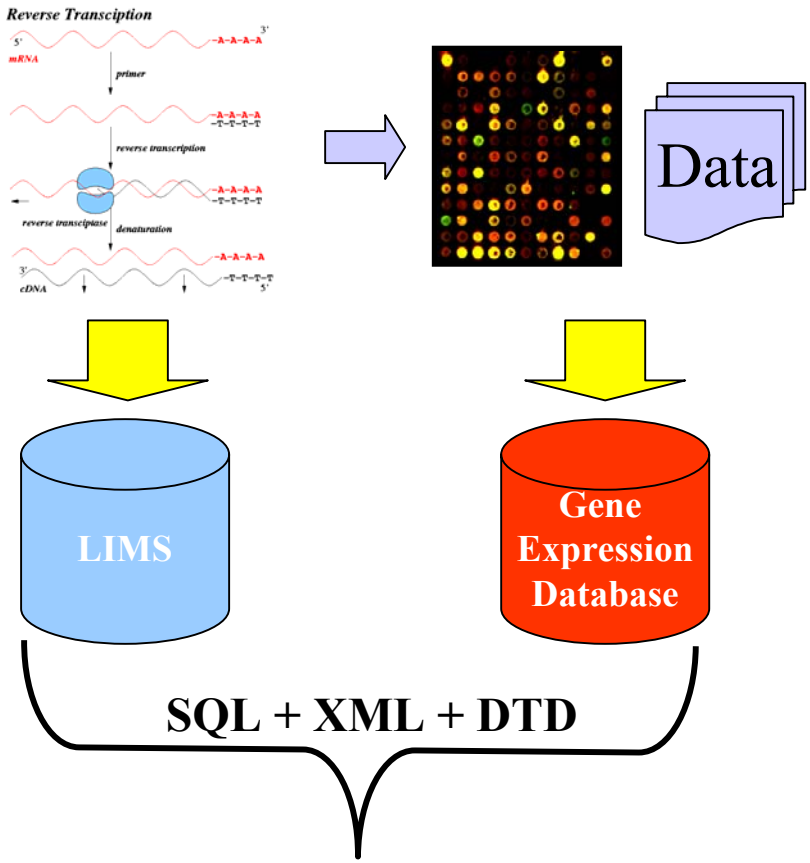
MAGE-stk (software tool kit)

http://www.mged.org/Workgroups/MAGE/magestk.html

How do you generate MAGE-ML ?

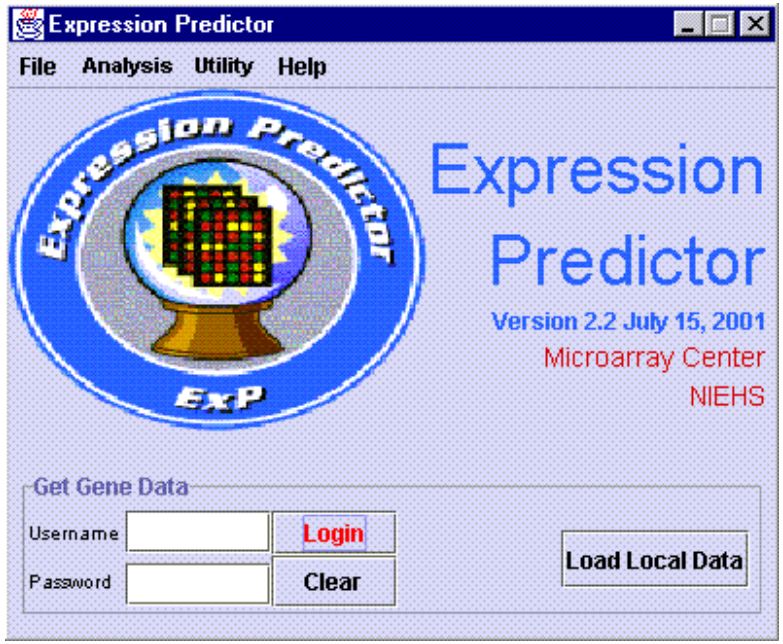
Generating MAGE-ML

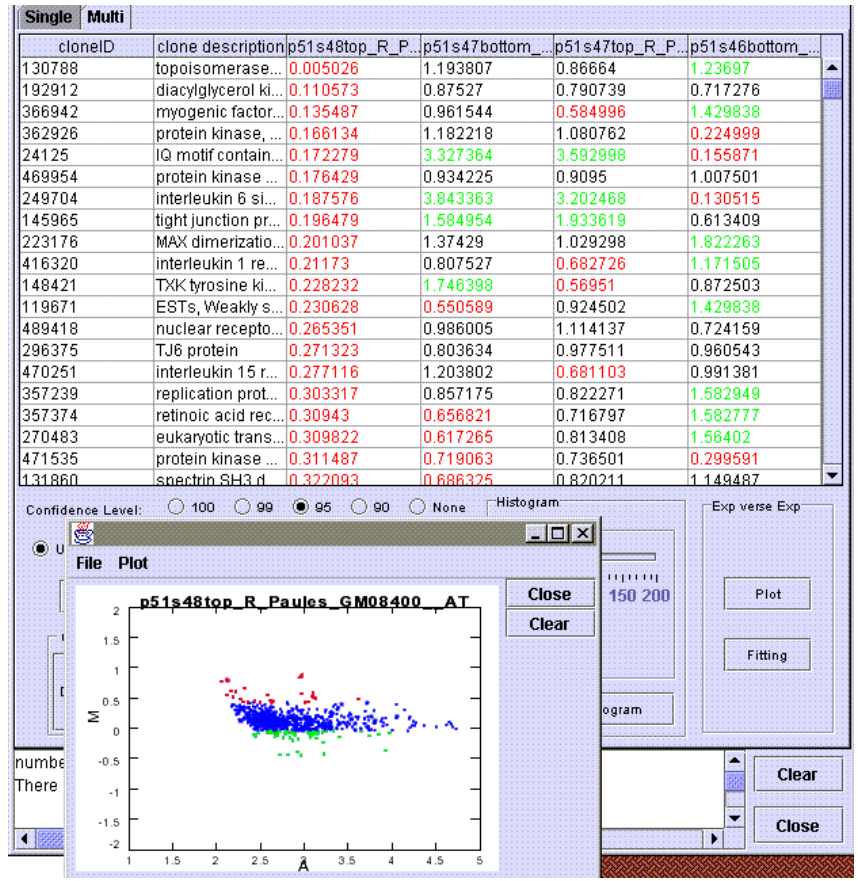
Route taken by Microarray Centers and Groups with adequate Bioinformatics Support



MAGE-ML information and data Document

Expression Predictor





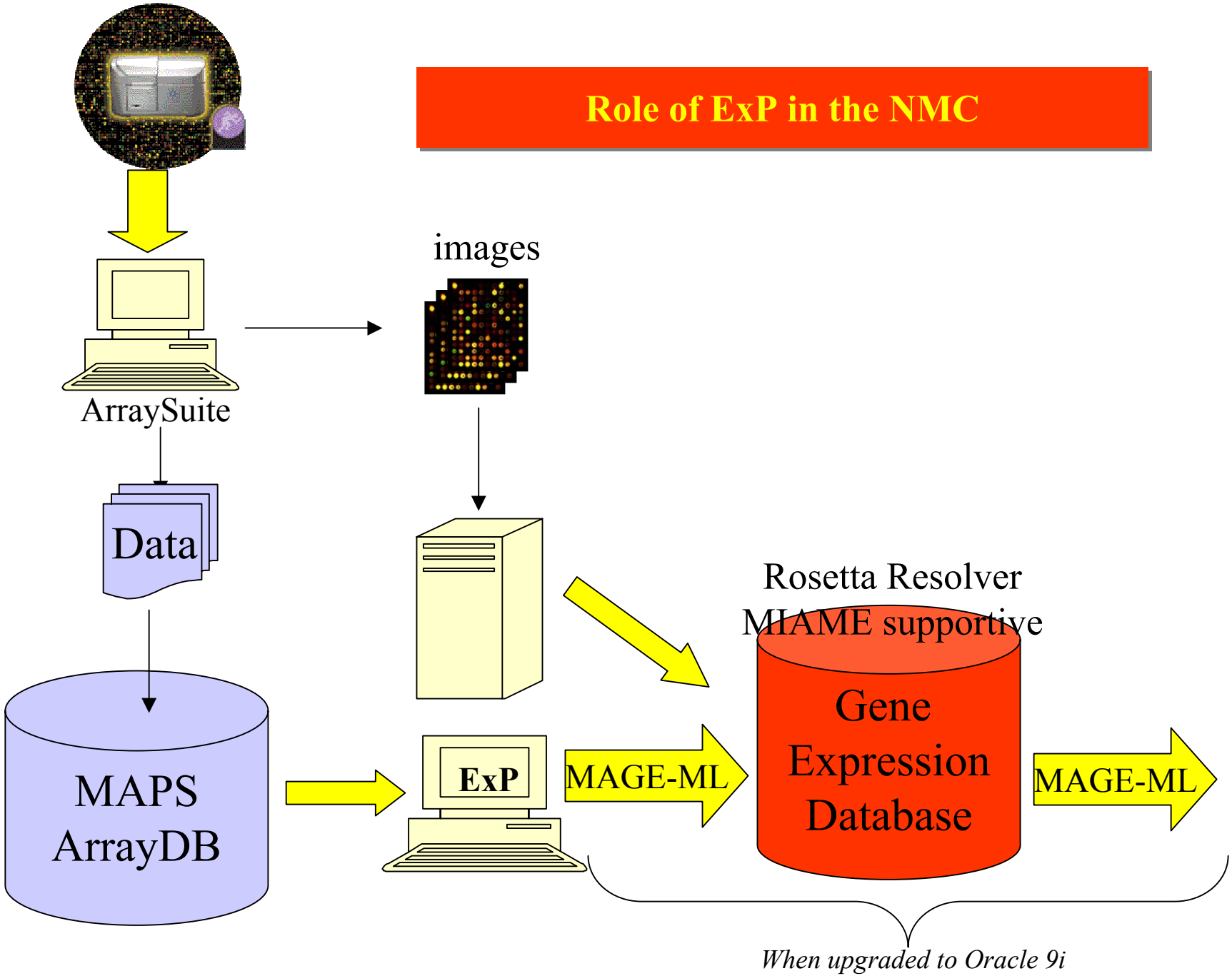
cloneID	clone description	p51s48top_R_P...	p51s47bottom_...	p51s47top_R_P...	p51s46bottom_...
130788	topoisomerase...	0.005026	1.193807	0.86664	1.23697
192912	diacylglycerol ki...	0.110573	0.87527	0.790739	0.717276
366942	myogenic factor...	0.135487	0.961544	0.584996	1.429838
362926	protein kinase, ...	0.166134	1.182218	1.080762	0.224999
24125	IQ motif contain...	0.172279	3.327364	3.592998	0.155871
469954	protein kinase ...	0.176429	0.934225	0.9095	1.007501
249704	interleukin 6 si...	0.187576	3.843363	3.202468	0.130515
145965	tight junction pr...	0.196479	1.584954	1.933619	0.613409
223176	MAX dimerizatio...	0.201037	1.37429	1.029298	1.822263
416320	interleukin 1 re...	0.21173	0.807527	0.682726	1.171505
148421	TXK tyrosine ki...	0.228232	1.746398	0.56951	0.872503
119671	ESTs, Weakly s...	0.230628	0.550589	0.924502	1.429838
489418	nuclear recepto...	0.265351	0.986005	1.114137	0.724159
296375	TJ6 protein	0.271323	0.803634	0.977511	0.960543
470251	interleukin 15 r...	0.277116	1.203802	0.681103	0.991381
357239	replication prot...	0.303317	0.857175	0.822271	1.582949
357374	retinoic acid rec...	0.30943	0.656821	0.716797	1.582777
270483	eukaryotic trans...	0.309822	0.617265	0.813408	1.56402
471535	protein kinase ...	0.311487	0.719063	0.736501	0.299591
131860	spectrin SH3 d...	0.322093	0.686325	0.820211	1.149487

ExP permits:

- Compiling gene expression data
- preprocessing data
- Evaluation of data
- Dimension reduction
- Classification and Prediction
- Export of data in microarray XML formats (GEML and MAGE-ML)

ExP is currently used by the NMG and in limited beta testing.
An application note, research paper, documentation and technology transfer supporting material are being put together for open distribution

Role of ExP in the NMC

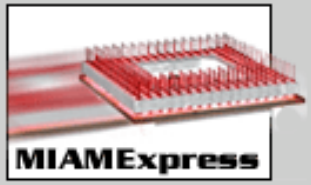


MIAME Software

Route taken by a PI or small lab with little bioinformatics support

- MIAMExpress - a MIAME annotation tool under development at the EBI.
- Gene Traffic - a microarray data management and analysis software developed at iobion informatics.
- Ipsogen Cancer Profiler - "A Bioinformatics system composed of Discovery Software tools and of ELOGE database, utilised to identify transcriptional signatures belonging to each cancer type. This system ensures the management of gene expression measurement, traceability in a GLP environment".
- BASE (BioArray Software Environment) - "a web-based open source MIAME-supportive microarray DB and analysis platform."
- GeneData Expressionist™ - "is a computational system that efficiently processes gene expression data generated by high-throughput microarray technologies. In support of MIAME Expressionist provides extensive reporting capabilities for data pre-processing, quality control and statistical analysis. The system is fully extendable to encompass all areas of MIAME".

http://www.mged.org/Workgroups/MIAME/miame_software.html



MIAMExpress Login
MIAMExpress Help
Contact Curation Team

Login: **test**

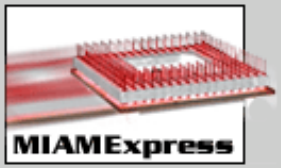
? Select type of submission

Protocol submission

Array design submission

Experiment submission

Browse existing array designs



MIAMExpress Login
MIAMExpress Help
Contact Curation Team

Login: **test**

Experimental Design:

? ID:

? Experiment type:

Normal vs. diseased comparison
Treated vs. untreated comparison
Time course
Dose response
Effect of gene knock-out
Effect of gene knock-in(transgenics)

If other, please specify:

? Experimental factors:

Age
Compound
Disease state
Cell line
Cell type
Developmental stage

If other, please specify:

? Experiment description:

What's the bottom-line to adhering to the microarray Standard ???

You can be either MIAME:

- Compliant – follow the MIAME and MAGE-ML specifications to the utmost

- Supportive – Capture the vital information and data about the microarray experiment

- Suggestive – Do what you can without compromising the data!

References

1: Adv Biochem Eng Biotechnol 2002;77:113-39

Microarray data representation, annotation and storage.

Brazma A, Sarkans U, Robinson A, Vilo J, Vingron M, Hoheisel J, Fellenberg K.

2: Genome Biol 2002 Aug 23;3(9):RESEARCH0046

Design and implementation of microarray gene expression markup language (MAGE-ML).

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A.