

caArray – cancer array informatics

Data Management:

- caArray database

Analysis Tools:

- caWorkbench
- webCGH

caArray Data Portal

The NCIC's cancer array informatics project, caArray, consists of a microarray database and microarray data analysis and visualization tools. CaArray is an open source project, and the source code and APIs are available in the [caArray Informatics](#) page. The goals of the project are to make microarray data publicly available, and to develop and bring together open source tools to analyze these data.

DevDoc
- Documentation for the open source caArray data repository, data portal servlet, and analysis tools

Data Management
- caArray database: A standards based repository for microarray data.

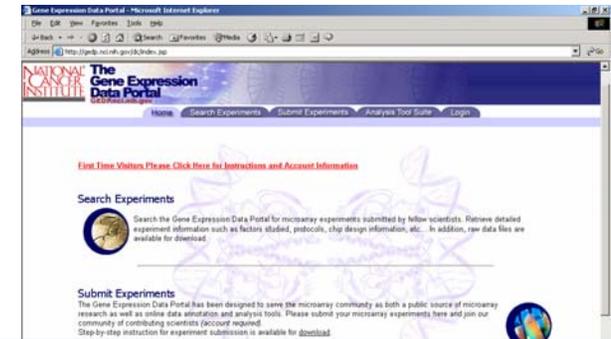
Data Analysis
- Data preprocessing, analysis and visualization tools.

<http://caarray.nci.nih.gov/>

Data Management

Gene Expression Data Portal (NCICB)

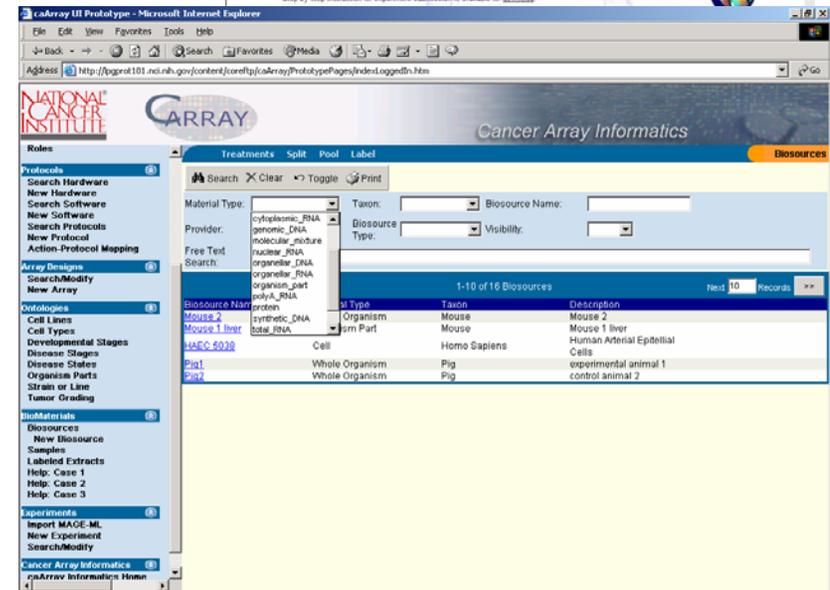
- Based on MIAME 1.0
- Public data repository
- Affymetrix, Spotted Array, CGH, and SAGE data



Available in September 2004:

caArray database (NCICB)

- Based on MIAME 1.1
- Public data repository
- Local deployment at cancer centers



Available now:

MIAMExpress (EBI)

- MAGE-ML import to caArray database

Survey of Existing MicroArray Systems

- ▶ 19 Existing MicroArray systems compared on the following criteria
 - Computationally MIAME 1.1 compliant
 - MAGE-ML Import
 - MAGE-ML Export
 - Affymetrix native file upload support
 - Genepix native file upload support
 - Security and data sharing capabilities
 - Ability to modify data after submission
 - Add/Remove Hybridizations
 - Correct mistakes in annotations
 - Architecture/Open source
 - **Ease of integration with caCORE technologies**
- ▶ No existing system was found to meet these needs

caArray Goals

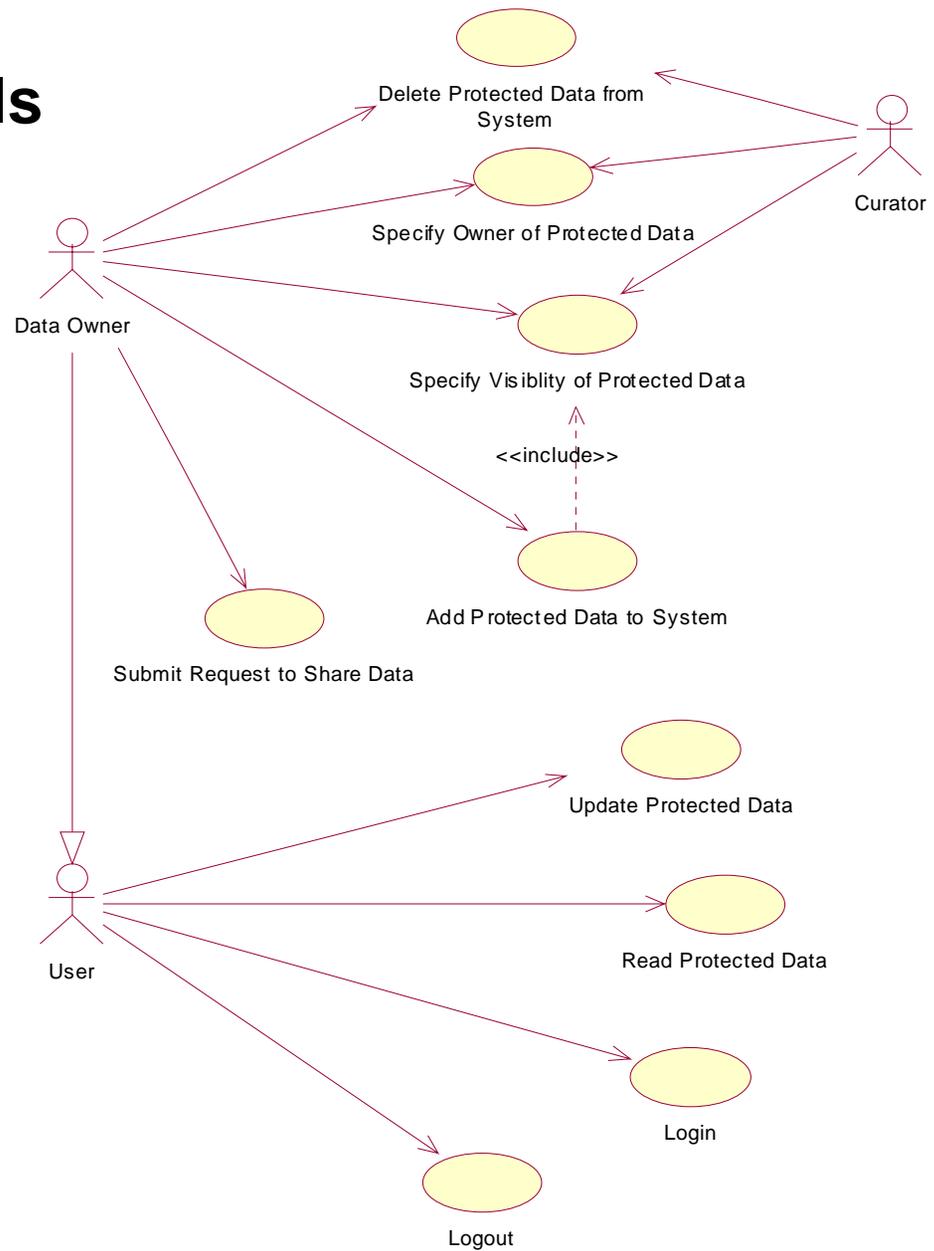
- ▶ Simplified data submission and annotation
 - Allow MAGE-ML import
 - Allow Affymetrix and GenePix native file uploads for Hybridizations
 - Provide MIAME 1.1 level annotations with intuitive interface
 - Controlled vocabularies (MGED ontology)
- ▶ Allow modification of data after original submission
- ▶ Interoperability
 - MAGE-ML import and export
 - Open API
- ▶ Leverage existing open source technologies
 - Utilizes the MAGE-stk Java toolkit
 - Uses Apache/Jakarta Object Relational Bridge to map MAGE objects to relational database
- ▶ Advanced data sharing capabilities

caArray Data Management Features

- ▶ Security
 - Supports sharing of data with arbitrary groupings of users
 - Supports public/private data
- ▶ Data Submission
 - MAGE-ML
 - Affymetrix and GenePix files
- ▶ Data Modification
 - All data can be modified by user with sufficient privileges after submission
 - Add/Remove Hybridizations
- ▶ Simplified MIAME 1.1 Annotations
 - Re-usable data definitions
 - Protocols, Hardware, and Software
 - Array designs
 - Biomaterials
 - Experiment
- ▶ BioMaterial Annotation
 - Create BioSource and annotate it
 - Apply treatments to derive biomaterials and labeled extracts

ManageDataAccessDetails

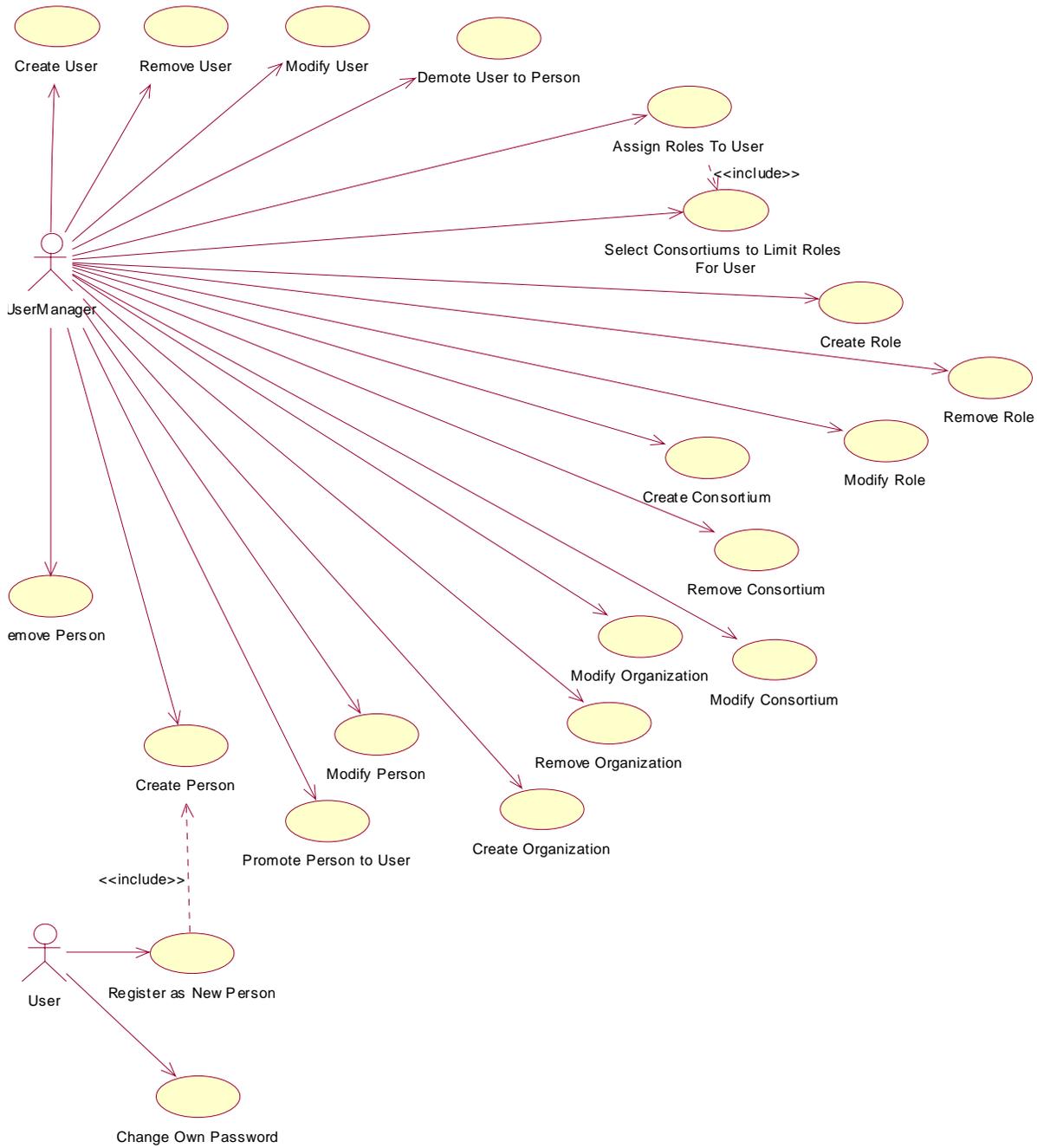
Data management
in the caArray system



caArray SoDA

Use Case Template:

<http://caarray.nci.nih.gov/caARRAY/devdoc/caarraydbdocs>



caArray UI prototype

<http://caarray.nci.nih.gov/caARRAY/devdoc/caarraydbdocs>

caArray UI Prototype - Microsoft Internet Explorer

File Edit View Favorites Tools Help

NATIONAL CANCER INSTITUTE CARRAY Cancer Array Informatics

caArray Data Management

caArray Data Portal

The NCICB's cancer array informatics project, caArray, consists of a microarray database and microarray data analysis and visualization tools. CaArray is an open source project, and the source code and APIs are available in the [caArray Informatics](#) page. The goals of the project are to make microarray data publicly available, and to develop and bring together open source tools to analyze these data.

DevDoc
- Documentation for the open source caArray data repository, data portal servlet, and analysis tools

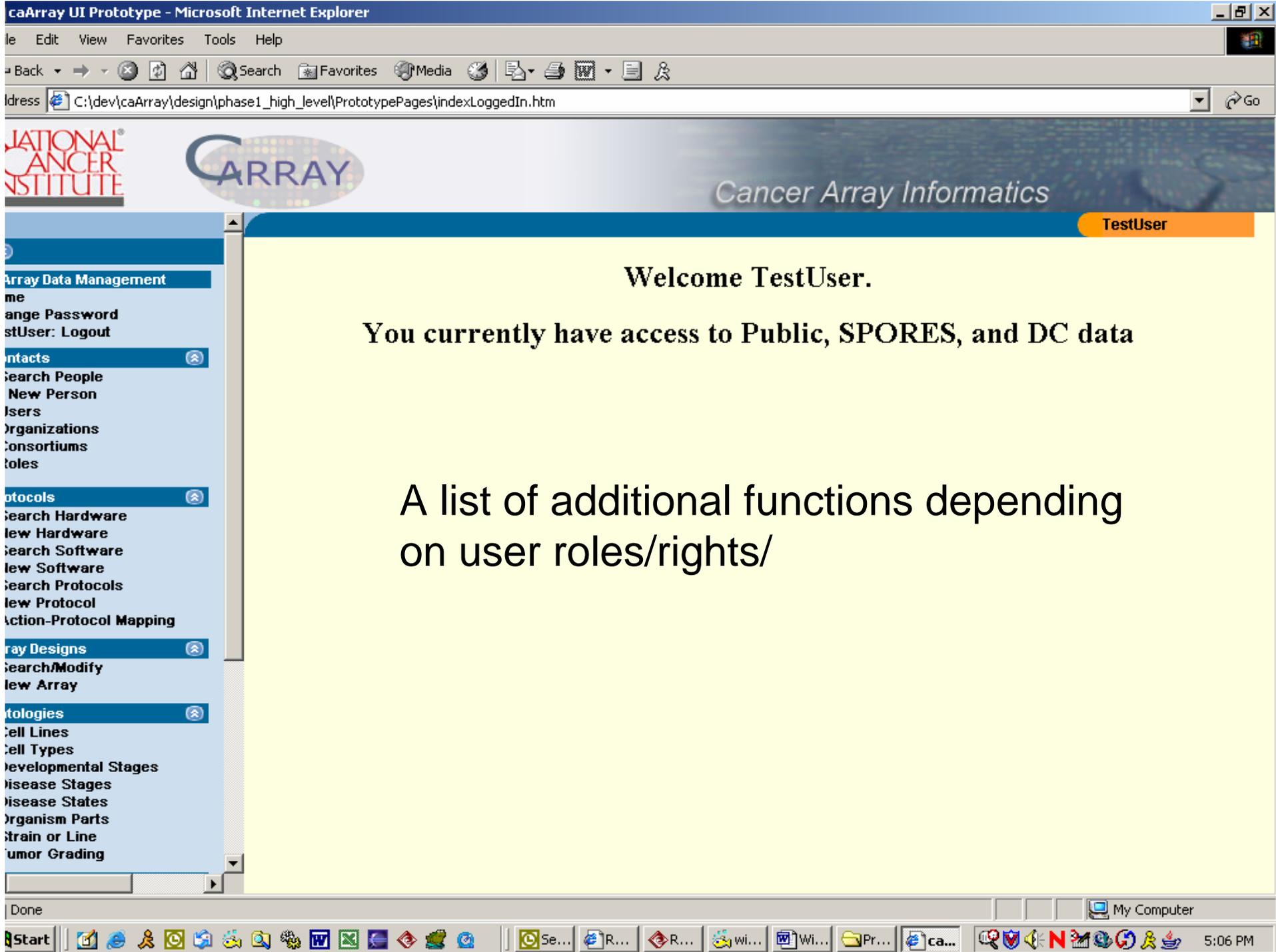
Data Management
- caArray database:
A standards based repository for microarray data.

Data Analysis
- Data preprocessing, analysis and visualization tools.

caArray Data Management
Home
Login
Register For An Account
180 Experiments
1250 Hybridizations
40 Array Designs
128 Protocols
Search Experiments
Search Array Designs
Search Protocols
Cancer Array Informatics
caArray Informatics Home

A link back to caArray Informatics

Start | I. | W. | p. | C. | T. | p. | W | u. | 4:07 PM



Welcome TestUser.

You currently have access to Public, SPORES, and DC data

A list of additional functions depending on user roles/rights/

Array Data Management

Change Password
Logout

Contacts

Search People
New Person
Users
Organizations
Consortiums
Roles

Protocols

Search Hardware
New Hardware
Search Software
New Software
Search Protocols
New Protocol
Action-Protocol Mapping

Array Designs

Search/Modify
New Array

Biologies

Cell Lines
Cell Types
Developmental Stages
Disease Stages
Disease States
Organism Parts
Strain or Line
Tumor Grading



Search Hardware

Search Clear Toggle Print

Hardware Type: Hardware Model:

Hardware Make: Hardware Manufacture:

Free Text Search:

Visibility:

1-10 of 16 Hardware

Next 10 Records >>

Hardware Model	Hardware Type	Hardware Make	Hardware Manufacture
Hardware Model1	Array Scanner	Hardware Make 1	Hardware Manufacture 1
Hardware Model2	Array Scanner	Hardware Make 2	Hardware Manufacture 2
Fluidics Station 450	hybridization_station	Affymetrix	Affymetrix, Inc.
Standard HybChamber	hybridization_chamber	GeneMachines	GeneMachines
Themomix 1480	waterbath	B. Braun	Westshore Technologies, Inc.
GenePix@ 4000E	array_scanner	Axon	Axon Instruments, Inc
Affymetrix 418 scanner	array_scanner	Affymetrix	Affymetrix, Inc.

Protocol Management

Dynamic Ontologies

The screenshot displays the 'caArray UI Prototype' in Microsoft Internet Explorer. The main page features the National Cancer Institute logo and 'CARRAY Cancer Array Informatics'. A search bar is visible with the following fields:

- Cell Line Name: [Empty]
- Public Database (for reference): All
- Visibility: All
- Cell Line Accession Number: [Empty]

Below the search bar, a table lists 1-10 of 16 Cell Lines. The first entry is:

Cell Line Name	Database	Accession Number	Link
2E10-H2	ACC 178	DSMZ	Link to ACC 178 in DSMZ database

An inset window titled 'Query Results - Microsoft Internet Explorer' shows the results for the query: `http://www.cabri.org/CABRI/srs-bin/wgetz?-newId+-e+-page+qResult+[DSMZ_MUTZ-id:'ACC%20178']`. It includes options to apply to selected or unselected results, result options (Link, Save), and display options. The results are as follows:

Accession_number	ACC 178
Cell_line_name	2E10-H2
Brief_description	confirmed as mouse with IEF of AST, NP established by fusion of SP2/0-AG14 myeloma cells with BALB/c mouse spleen cells immunized with purified canine adenovirus type 1 (CAV-1); the secreted antibody binds to and neutralizes CAV-1 and is type-specific
Description	mouse hybridoma (anti-canine adenovirus type 1) established by fusion of SP2/0-AG14 myeloma cells with BALB/c mouse spleen cells immunized with purified canine adenovirus type 1 (CAV-1); the secreted antibody binds to and neutralizes CAV-1 and is type-specific

archy

- [-] [C] BioMaterial #1
- [-] [C] BioMaterialCharacteristics #1
 - [C] Age #1
 - [C] BioSourceProvider #1
 - [C] BioSourceType #1
 - [C] Biometrics #1
 - [C] CellLine #1
 - [C] CellType #1
 - [C] ChromosomalAberrationClassification #1
 - [C] ClinicalTest #1
 - [C] ClinicalTestType #1
 - [C] ClinicalTreatment #1
 - [C] DevelopmentalStage #1
 - [C] DiseaseStaging #1
 - [C] DiseaseState #1
 - [-] [C] EnvironmentalHistory #1
 - [C] Bedding #1
 - [C] ClinicalHistory #1
 - [C] FamilyHistory #1
 - [C] Generations #1
 - [C] GeographicLocation #1
 - [-] [C] GrowthCondition #1
 - [C] Atmosphere #1
 - [C] BarrierFacility #1
 - [C] Humidity #1
 - [C] Light #1
 - [C] Media #1
 - [C] Nutrients #1

Hierarchy

- [C] Nutrients #1
- [C] PopulationDensity #1
- [C] Temperature #1
- [C] Water #1
- [C] Host #1
- [C] PathogenTests #1
- [C] GeneticModification #1
- [C] Histology #1
- [C] Individual #1
- [-] [C] IndividualGeneticCharacteristics #1
 - [C] Allele #1
 - [C] ChromosomalAberration #1
 - [C] Genotype #1
 - [C] Haplotype #1
 - [C] IndividualChromosomalAbnormality #1
 - [C] Ploidy #1
- [C] Organism #1
- [-] [C] OrganismPart #1
 - [C] DiseaseLocation #1
- [C] OrganismStatus #1
- [C] Phenotype #1
- [C] Sex #1
- [-] [C] StrainOrLine #1
 - [C] Cultivar #1
 - [C] Ecotype #1
- [C] TargetedCellType #1
- [C] TumorGrading #1
- [C] BioSampleType #1

ers

BioMaterialCharacteristics #1

Supers

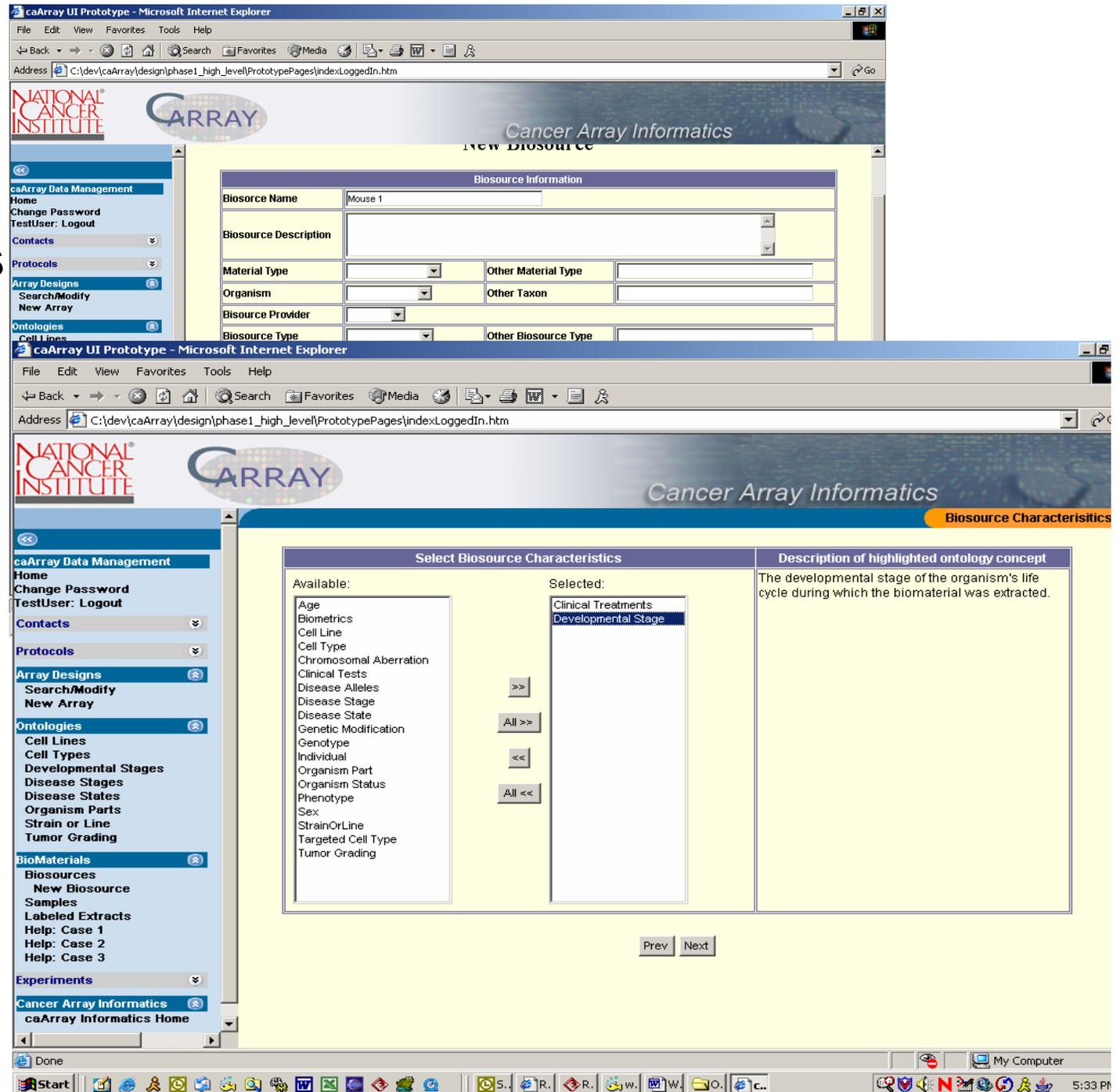
[C] BioMaterialCharacteristics #1

BioMaterialCharacteristics

Done

BioMaterial Characteristics

Specify relevant characteristics



caArray UI Prototype - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\dev\caArray\design\phase1_high_level\PrototypePages\index.LoggedIn.htm

NATIONAL CANCER INSTITUTE **CARRAY** *Cancer Array Informatics*

Biosource: Mouse 2

Transfer Ownership Duplicate Delete Done

Biosource Information	
Biosource Name	Mouse 2
Biosource Description	Some description
Material Type	whole_organism
Organism	Mouse
Biosource Provider	Joe Doe
Biosource Type	paraffin_sample

Modify

Visibility	
Visibility	Public

Modify

Biosource Characteristics			
Age:	<input type="text"/> years	Initial Time Point:	<input type="text"/>
Min Age:	<input type="text"/> years	Max Age:	<input type="text"/> years
Biometrics:	Weight <input type="text"/> lbs	Height <input type="text"/> cm	Other: <input type="text"/>
Cell Line	<input type="text"/>		
Cell Type	<input type="text"/>		
Chromosomal	Type(s): <input type="text"/>		

gedIn.htm

Cancer Array Informatics

Disease State	<input type="text"/>	
Genetic Modification	Type: <input type="text"/>	Description: <input type="text"/>
Genotype	<input type="text"/>	
Histology	<input type="text"/>	
Individual ID	<input type="text"/>	
Organism Part	<input type="text"/>	
Organism Status	<input type="text"/>	
Phenotype	<input type="text"/>	
Ploidy	<input type="text"/>	
Sex	<input type="text"/>	
Strain or Line	<input type="text"/>	
Targeted Cell Type	<input type="text"/>	
Tumor Grading	<input type="text"/>	
Additional Description:	<input type="text"/>	

Save Prev Cancel

Done

My Computer

5:36 PM

Start

Contacts

Protocols

Array Designs

Search/Modify

New Array

Ontologies

Cell Lines

Cell Types

Developmental Stages

Disease Stages

Disease States

Organism Parts

Strain or Line

Tumor Grading

BioMaterials

Biosources

New Biosource

Samples

Labeled Extracts

Help: Case 1

Help: Case 2

Help: Case 3

Experiments

Cancer Array Informatics

caArray Informatics Home

5:37 PM

Start

Done

My Computer

- caArray Data Management
- Home
- Change Password
- TestUser: Logout
- Contacts
- Protocols
- Array Designs
- BioMaterials
- Biosources
 - New Biosource
 - Samples
 - Labeled Extracts
 - Help: Case 1
 - Help: Case 2
 - Help: Case 3
- Experiments
- Cancer Array Informatics

Biomaterial Treatment (biosample creation).

Biomaterial Information																
Biomaterial creation description:	<input type="text" value="Mouse 2 Pancreas RNA extraction"/>															
Select source biomaterial	<table border="0"><tr><td>Available:</td><td></td><td>Selected:</td></tr><tr><td>HAEC 5038</td><td>>></td><td>Mouse 1 liver</td></tr><tr><td>Pig 1</td><td>All >></td><td>Mouse 2</td></tr><tr><td>Pig 2</td><td><<</td><td></td></tr><tr><td></td><td>All <<</td><td></td></tr></table>	Available:		Selected:	HAEC 5038	>>	Mouse 1 liver	Pig 1	All >>	Mouse 2	Pig 2	<<			All <<	
Available:		Selected:														
HAEC 5038	>>	Mouse 1 liver														
Pig 1	All >>	Mouse 2														
Pig 2	<<															
	All <<															
Name resulting biomaterial	<input type="text" value="Mouse 2 Pancreas RNA"/>															
Type of resulting biomaterial	<input type="text" value="total_RNA"/>															
Description of resulting biomaterial	<input type="text" value="Pancreas dissected from Mouse 2; total RNA extracted from pancreas"/>															

Next Cancel

caArray UI Prototype - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\dev\caArray\design\phase1_high_level\PrototypePages\index.htm

NATIONAL CANCER INSTITUTE **CARRAY** *Cancer Array Informatics*

General Experiment Information

Title:	Experiment1
Experiment Date:	Date of completion of the experiment
Experiment Type	Normal vs. Diseased Comparison
Visibility	Public
Description/Abstract:	An overall general description of the purpose of the experiment, experiment design, biosamples, extract preparation and labeling, hybridization procedures and parameters, and derived measurements

[Modify](#)

Contacts

Principal Investigator:	Scott Melby
Contact:	Juergen Lorenz

[Modify](#)

Experimental Factors/Variables

Factor	Factor Type	Scale	Levels	Actions
Age	Type 1	ordinal	1-12 Months 13-24 Months <input type="text"/>	Remove
Cell type	Type 2	nominal	Cell type1 Cell type2 <input type="text"/>	

[Add A New Factor](#)

New Experiment: General Information and Experimental Factors

File Upload

caArray UI Prototype - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\dev\caArray\design\phase1_high_level\PrototypePages\index.htm

NATIONAL CANCER INSTITUTE **CARRAY** *Cancer Array Informatics*

Image Acquisition Protocol:	Affymetrix 418 scanning
Feature Extraction Protocol:	Affymetrix MAS 5.0 Absolute Expression Analysis

Hybridizations (File Uploads)

#	Array Batch	Array Identifier	Labeled Extracts	Factor Levels	Files to upload
1	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 1-12 Months Cell type: Cell type 1	.cel <input type="text"/> Browse... .exp <input type="text"/> Browse... .bt <input type="text"/> Browse...
2	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 1-12 Months Cell type: Cell type 2	.cel <input type="text"/> Browse... .exp <input type="text"/> Browse... .bt <input type="text"/> Browse...
3	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 13-24 Months Cell type: Cell type 1	.cel <input type="text"/> Browse... .exp <input type="text"/> Browse... .bt <input type="text"/> Browse...
4	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 13-24 Months Cell type: Cell type 2	.cel <input type="text"/> Browse... .exp <input type="text"/> Browse... .bt <input type="text"/> Browse...

[Upload](#)

Done

Start | ScreenShots... | PrototypePa... | caArray UI ... | NetWare Me... | 6:23 PM

Upload
processed data files

caArray UI Prototype - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: C:\dev\caArray\design\phase1_high_level\PrototypePages\indexLoggedIn.htm

NATIONAL CANCER INSTITUTE ARRAY Cancer Array Informatics

Additional Data Processing

caArray Data Management

Home

Change Password

TestUser: Logout

Contacts

Protocols

Array Designs

BioMaterials

Experiments

Import MAGE-ML

New Experiment

Search/Modify

Cancer Array Informatics

Additional Data Processing

File Name: Browse...

Apply Protocol: RMAExpress qualification

URI: http://stat-www.berkeley.edu/~bolstad/RMAExpress/RMAExpress.html

Protocol Description: RMA is the Robust Multichip Average. It consists of three steps: a background adjustment (optional), quartile normalization(optional), and finally summarization.

Software Description: RMAExpress is a standalone GUI program for Windows (and Linux) to compute gene expression summary values for Affymetrix Genechip data using the Robust Multichip Average expression summary. It does not require R nor is it dependent on any component of the Bioconductor project.

Name	Value	Type
normalization	<input type="text"/>	String
background adjust	<input type="text"/>	String

Parameters:

Upload Cancel

Upload
any other files

Done

Start

WireFrameScr... PrototypePages caArray UI P... 7:19 PM

caArray Data Management

Home

Change Password

TestUser: Logout

Contacts

Protocols

Array Designs

BioMaterials

Experiments

Import MAGE-ML

New Experiment

Search/Modify

Cancer Array Informatics

Upload Files

File Name: Browse...

Description: The description for the file 1

Upload Cancel

Done

Start

WireFrameScr... PrototypePages caArray UI P... 7:19 PM

Data Curation

1. Information submitted via caArray user interface:

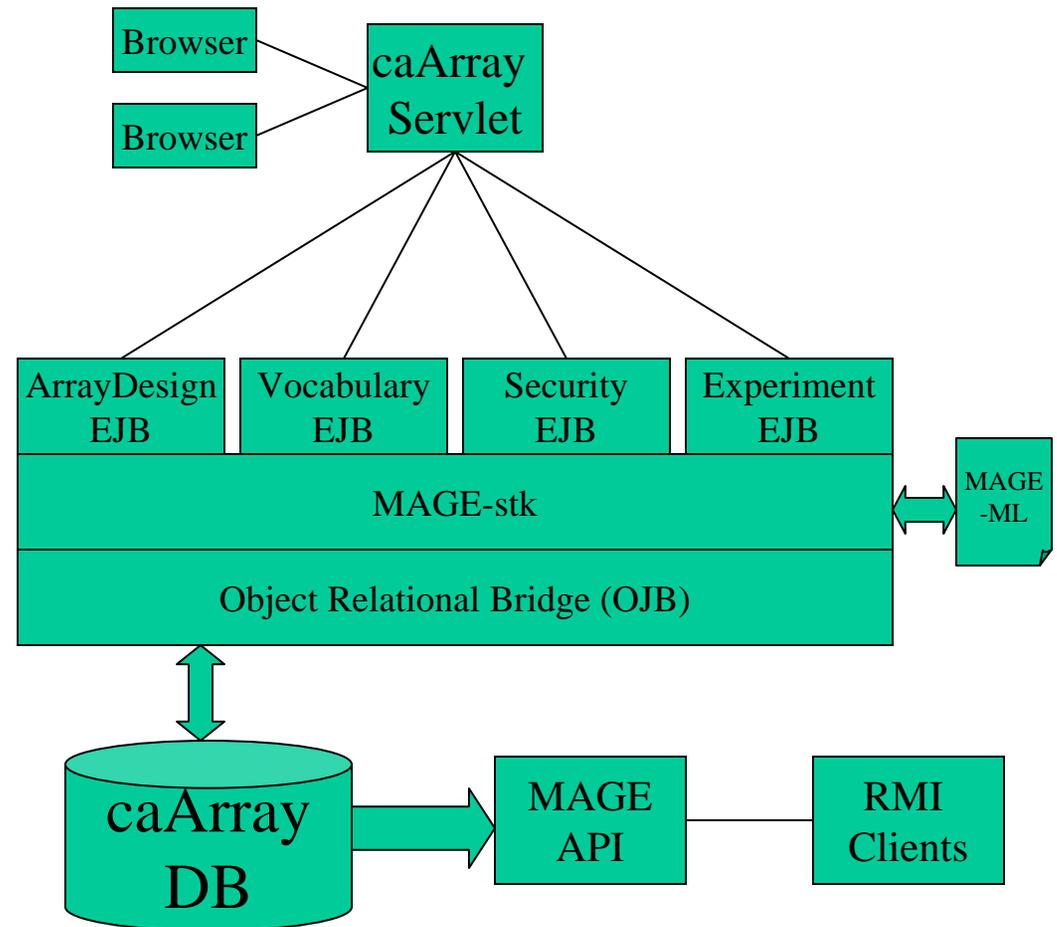
- Local caArray installation:
Each center can have its own data curation policies.
- Data annotation according to the published MIAME checklist.

2. MAGE-ML curation:

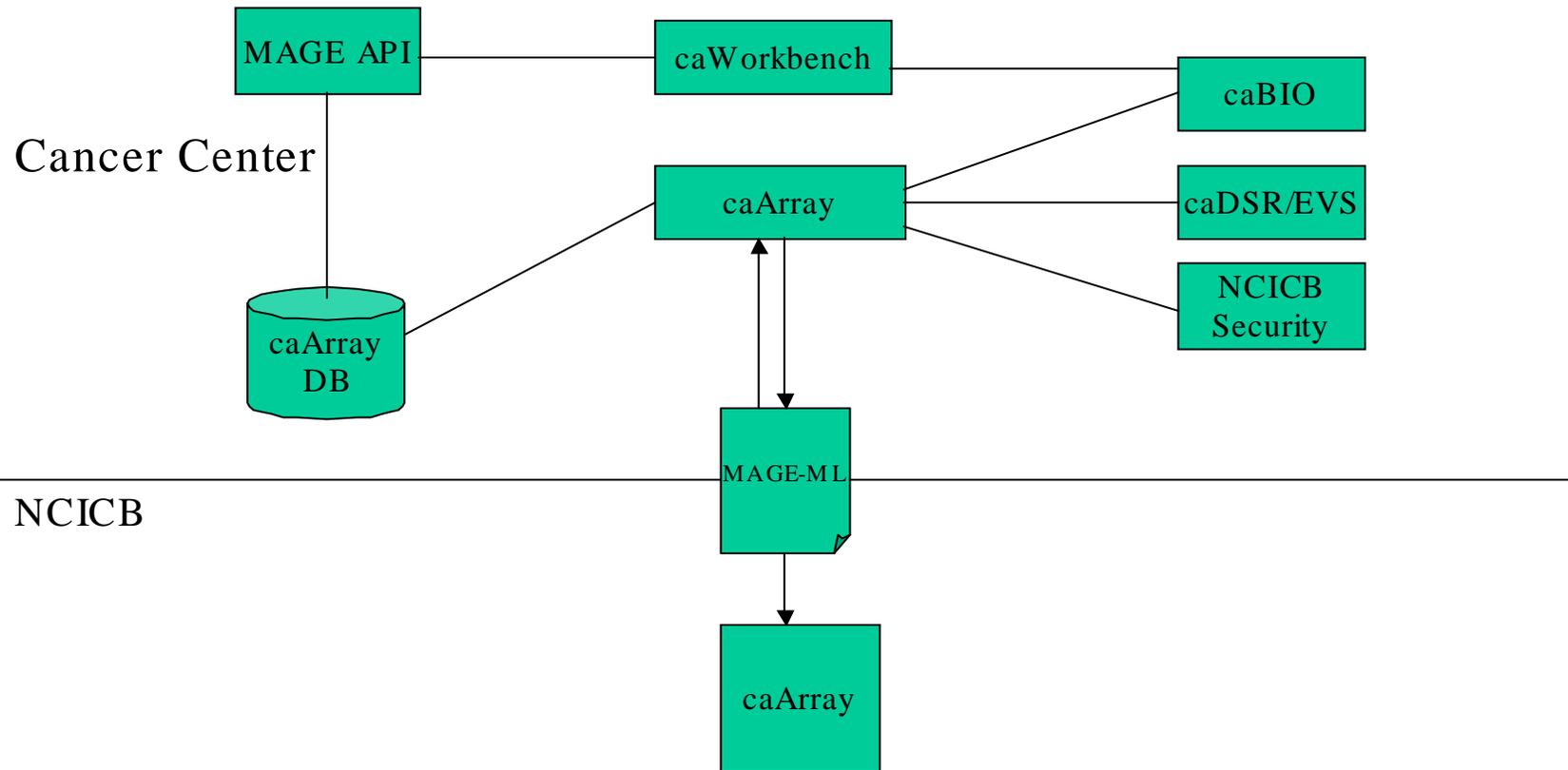
- MAGE-ML exchange between different caArray installations does not require additional curation.
- MAGE-ML exchange between caArray and other systems (e.g. GeneTraffic) may require curation.

caArray Architecture

- ▶ N-tier architecture
 - Struts/Servlet based user interface
 - EJB API layer to access services
 - Data Access Objects (DAO) encapsulation of persistence layer
 - Object Relational Bridge (OBJ) used to map MAGE-stk objects to relational database
 - MAGE-stk used for MAGE-ML import/export
- ▶ MAGE API provides RMI read only access to data

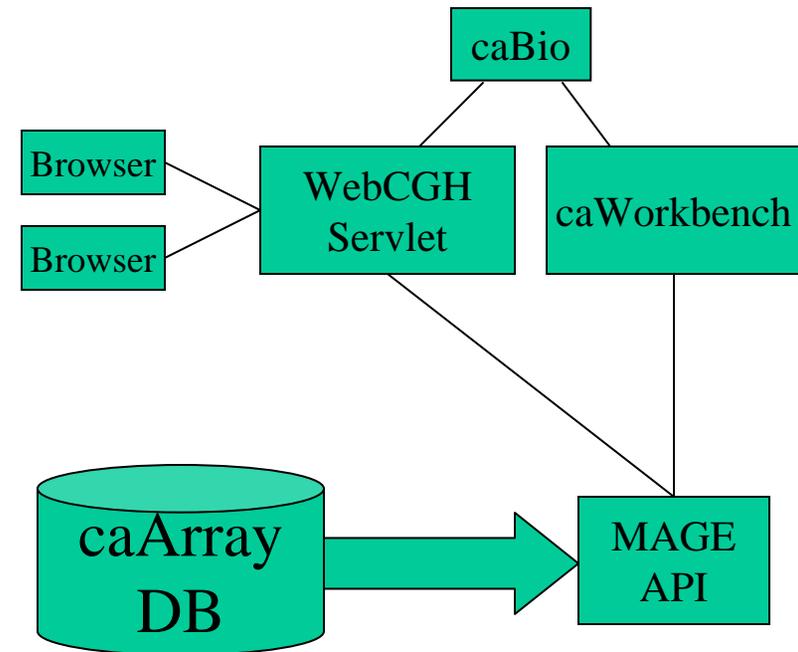


caArray System at Cancer Centers



caArray Analysis Tools

- ▶ WebCGH
 - Struts/Servlet application
 - Utilizes MAGE-API RMI interface to obtain data from caArray database
- ▶ caWorkbench
 - Java Swing application
 - Utilizes MAGE-API RMI interface to obtain data from caArray database

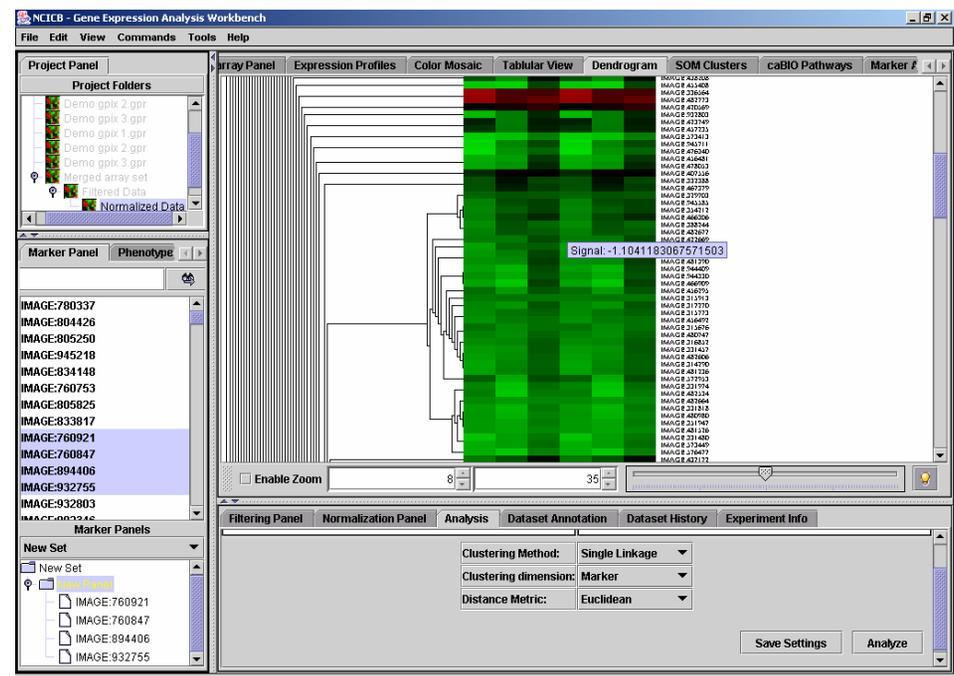


caWorkbench

- A suite of analysis, visualization and annotation functions for microarray data.
- Allows access (via the caBio API) to data sources containing information relating to genes and pathways.
- The long term goal is to evolve caWorkbench into a flexible, configurable and integrated platform that provides access to the data and annotations hosted by NCI and affiliated institutions.

caWorkbench v1.0:

- ▶ caWorkbench v1.0 released in September 2003
- ▶ Application is available at <http://ncicb.nci.nih.gov/download/>
- ▶ V1.0 features:
 - Color mosaic view
 - Microarray view
 - Tabular view
 - Data filtering
 - Data normalization
 - Hierarchical clustering
 - Self Organizing Maps
 - BioCarta pathways
 - CGAP gene annotations



caWorkbench v2.0

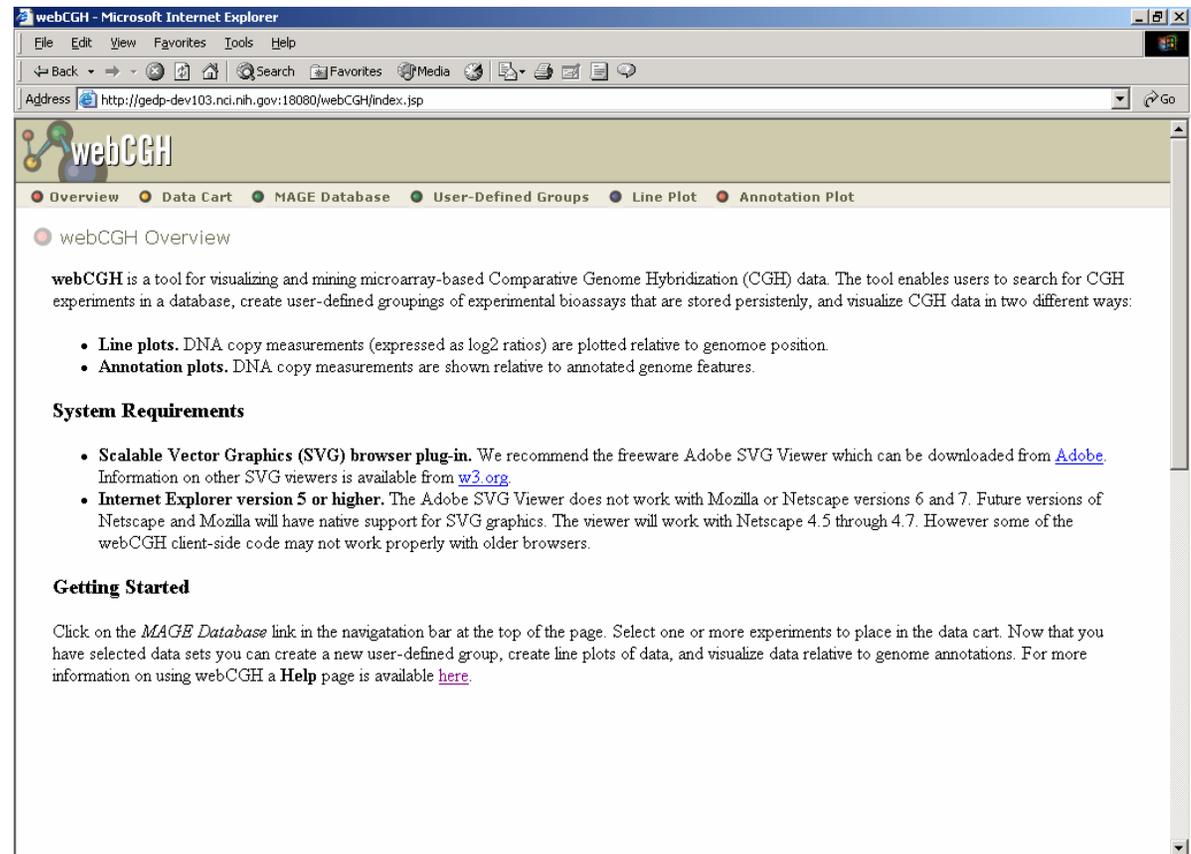
- **Performance and usability enhancements**
 - multiple color schemes
 - caching of caBio calls
 - analysis workflow builder
- **Support for new data types**
 - UCSF SPOT, Array Suite, proteomics
- **New analysis components:**
 - scatter plots, whole genome plots, value distribution, principal component analysis, additional normalizers and filters.
- **Annotations and via caBio:**
 - UCSC genome annotations, CGAP data
- **Data Integration**
 - clinical, calmage

webCGH

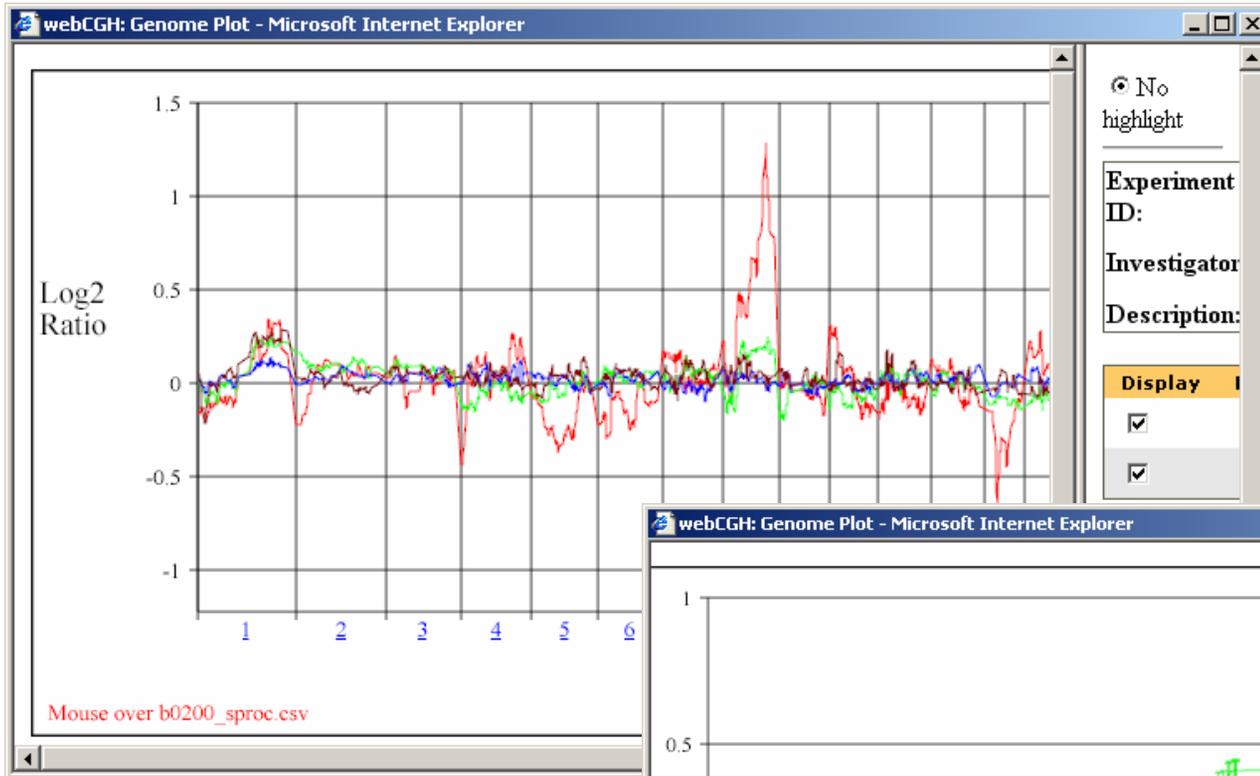
Data visualization and mining tool for array-based CGH data

- webCGH v1.0 available in March 2004

1. Line plots
2. Annotation plots



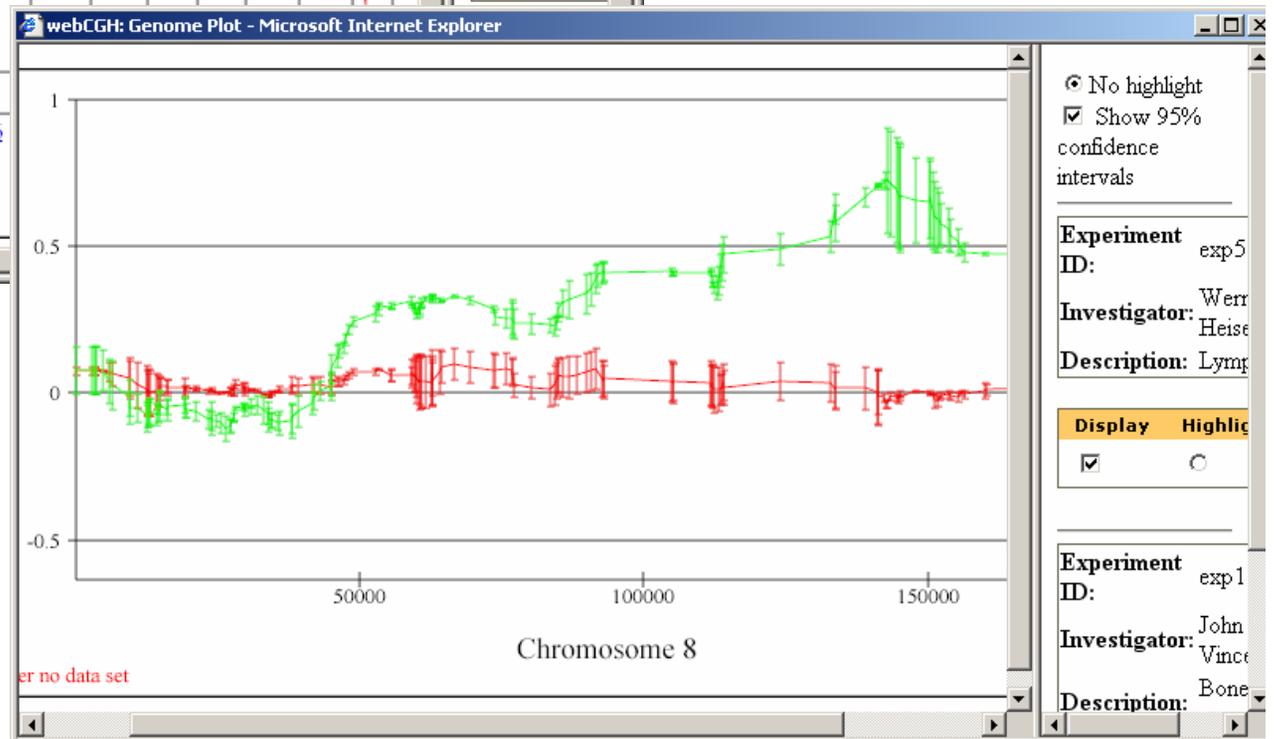
Line Plots



Genome Plot



Chromosome Plot



Annotation Plot

webCGH: Annotation Plot - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Genome Annotation Types

<input type="checkbox"/> acembly	<input type="checkbox"/> affyGeno
<input type="checkbox"/> all_fosends	<input type="checkbox"/> all_sts_primer
<input type="checkbox"/> blastzBestMm3	<input type="checkbox"/> blastzTightMm3
<input type="checkbox"/> est	<input type="checkbox"/> gl
<input type="checkbox"/> mouseChain	<input type="checkbox"/> mouseChainLink
<input type="checkbox"/> zoom2500_hq15Mm3L	<input type="checkbox"/> zoom50_hq15Mm3
<input type="checkbox"/> ensGene	<input type="checkbox"/> estOrientInfo
<input type="checkbox"/> fosEndPairs	<input type="checkbox"/> qcPercent
<input type="checkbox"/> genscan	<input type="checkbox"/> genscanSubopt
<input checked="" type="checkbox"/> knownGene	<input type="checkbox"/> mqcFullMrna
<input type="checkbox"/> mrnaOrientInfo	<input type="checkbox"/> multizMm3Rn2
<input type="checkbox"/> recombRate	<input type="checkbox"/> refGene
<input type="checkbox"/> snpGene	<input type="checkbox"/> snpNih
<input type="checkbox"/> stsMap	<input type="checkbox"/> syntenyMouse
<input type="checkbox"/> uniGene_2	<input type="checkbox"/> yeqaGene
<input type="checkbox"/> xenoMrna	

Plot width:

webCGH: Annotation Plot - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Go Chromosome: Start: KB End: KB Go

<<< << < Zoom In Zoom Out > >> >>>

Genome Position (kb)

146,000 146,200 146,400 146,600 146,800

CGH Fold Change (log2)

Fold change color code 0.70637 0.83996

2959

knownGene

AK092777

BC015181

AK024842

AF086559

AK021796

AK092777

BC048439

BC015181

AF086170

AK024842

BC010001

mrna

The plot displays a horizontal bar representing CGH fold change (log2) across a genomic region from 146,000 to 146,800 kb. The bar is color-coded from blue (low fold change) to red (high fold change). A legend indicates a color scale from 0.70637 (blue) to 0.83996 (red). The bar shows two distinct regions of low fold change (blue) with values 0.585 and 0.481. Below the bar, annotated genome features are listed, including known genes (AK092777, BC015181, AK024842) and mRNA transcripts (AF086559, AK021796, AK092777, BC048439, BC015181, AF086170, AK024842, BC010001).

webCGH v2.0

- **Simultaneous visualization of gene copy number and gene expression data.**
 - Support for additional data formats.
 - Clone/probe annotation via caBIO.
 - Integration with cytogenetic databases.
 - Simultaneous visualization of human and mouse data.
 - Ability to save images and data.
 - Queries of specific genes across different experiments.
 - Select a group of relevant genes and generate a report.
 - Visualize data along chromosome ideograms.