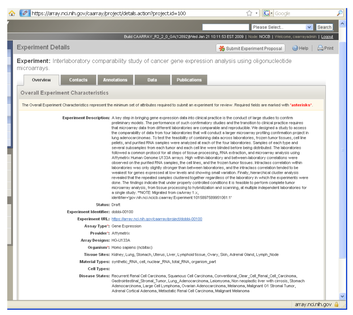
Data submission and curation for caArray, a standard based microarray data repository at NCI CBIIT

Bian X[1], Klemm J[1], Basu A[1], Hadfield J[1], Srinivasa R[2], Parnell T[2], Miller S[2], Mason W[2], Kokotov D[2], Duncan M[2], Duvall P[3], Gurses L[3], Boal T[4], Misquitta L[4], Swan D[5], Wysong R[5], Klink A[5], Johnson A[5], Fontenay G[6], Liu J[7], Colbert M[7], Komatsoulis G[1]
[1]NCI Center for Biomedical Informatics and Information Technology, [2]5AM Solutions, Inc., [3]Stelligent, Inc., [4]NARTec, Inc., [5]TerpSys, Inc., [6]Lawrence Berkeley National Laboratory , [7]Science Applications International Corporation (SAIC)
National Laboratory

Overview

caArray is an open-source, web and programmatically accessible array data management system. caArray guides the data submission with MAGE-TAB, a spreadsheet-based file format, which facilitates comprehensive annotation with standard ontology and terminology and easy to build. Careful curation of the data submitted ensures data in high quality and abide by community standard for easy data sharing and exchanging

Home Page



Experiment Overview Page

- **Store array data associated with experiment and sample annotations - data can be entered through the web interface or through MAGE-TAB**
- **Parse Affymetrix, Illumina and GenePix formats for expression and SNP array - store native files for other providers**
- **Data security and access control**
- **Manage protocols and controlled vocabularies**
- **Basic Browse and Search Functionality**
- **Data files and annotation download**
- **Programmatic access via a Java API and grid service**

caArray and MAGE-TAB

Data can be loaded into caArray using MAGE-TAB, a spreadsheet-based format for annotating and communicating microarray data in a MIAME-compliant fashion - http://www.mged.org/mage-tab

MAGE-TAB Components:

➢Investigation Description Format (IDF): General information about the investigation including its name, a brief description, the investigator's contact details, bibliographic references, etc.

➢Sample and Data Relationship Format (SDRF): Describes the relationships between the samples, arrays, data, and other objects used or produced in the investigation.

➢Array Design Format (ADF): Describes the design of an array

Sample IDF

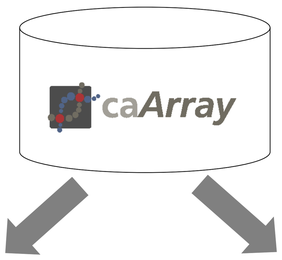
```
# This section contains the top-level information for your experiment.
Investigation Title      SNPs Genotyping for Cancer Cell Line Project
Experiment Description   Multiple genome-wide microarray 500K SNP studies performed to measure the genotype profile of various
Experimental Design      disease_state_design
Experimental Design Term Source REF      MO

# Please create as many Experimental Factors here as you need to
# describe the variables investigated by your experiment.
Experimental Factor Name        DISEASESTATE
Experimental Factor Type        disease_state
Experimental Factor Term Source REF      MO

"# Quality Control Type examples: dye_swap_quality_control, biological_replicate, technical_replicate"
Quality Control Type
Quality Control Term Source REF MO
```

Sample SDRF

# The first line below has been filled in with example terms, where possible:									
Source Name	Material Ty	Term Sourc	Characteri	Term Source REF	Characteristics[CellType]	Term Sour	Characteristics[Organism]	Characteri	Term Sou
BT474 - Replicate	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa
BT474 - Replicate	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa
BT474 - Replicate	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa
SKBR3 - Replica	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa
SKBR3 - Replica	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa
SKBR3 - Replica	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa

Integration with Data Analysis Tools

geWorkbench and GenePattern, bioinformatics platforms for molecular analysis, can pull data directly from caArray.

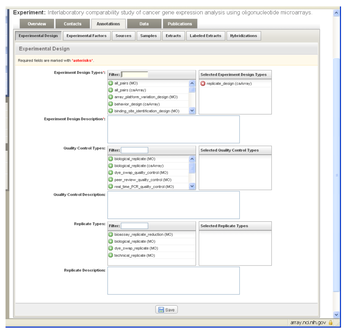


GenePattern
A platform for integrative genomics

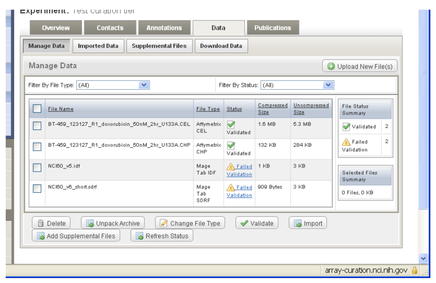
- **Designed to encourage the rapid integration of new techniques**
- **Current library includes over 100 analysis modules**
  - **Gene Expression Analysis**
  - **SNP Analysis**
  - **Data Conversion Modules**
- **Supports sharing of analysis workflows through its pipeline engine**

- **Visualize gene expression data in a variety of ways**
- **Access to client- and server-side analysis tools**
- **Validate computational hypotheses through the integration of gene and pathway information**

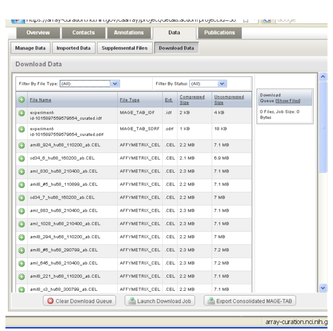
Data management



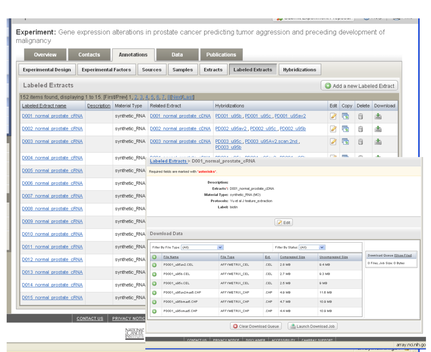
Annotation can be imported with MAGE-TAB files or entered through the portal



Data validation before import ensure proper format and abide by MAGE-TAB specification



User with proper privilege can download data of their choice and consolidated annotation



The relationship between biomaterials and their associated data files can be edited, viewed and downloaded

Data Curation

Purpose:
- **Data correctly parsed into and retrieved from the database**
- **Unambiguous presentation and report of research**
- **Easy comparison of results from different labs**
- **Compatible with other databases**
- **Easy data sharing and exchanging**

What we do:
- **Following MAGE-TAB specification**
- **Checking file formats**
- **Checking that data files match the array**
- **Checking biomaterials and data files are correctly associated**
- **Checking redundant use of common protocols**
- **Checking annotation quality**
- **Make sure annotation use controlled vocabulary and/or ontology**
- **Make sure annotation is correct and meaningful**
- **Help users in data submission, management and maintance**

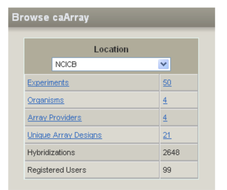
Supported Formats

File Type	*Parsed Formats*	*Native Files Only*
Array data	Affymetrix: .cel, .chp Illumina: .csv GenePix: .gpr	Affymetrix: .dat, .exp, .rpt, .txt Illumina: .idat, .txt Agilent: .tsv, .txt Nimblegen: .txt UCSF SPOT: .spt
Array designs	Affymetrix: .cdf Illumina: .csv GenePix: .gal	Agilent: .csv, .xml Imagene: .tpl Nimblegen: .ndf UCSF SPOT: .spt
MAGE-TAB	IDF SDRF	ADF Data matrix

caArray in the Community

Local Installations

caArray hosted at CBIIT



https://array.nci.nih.gov
Hosted data sets include:

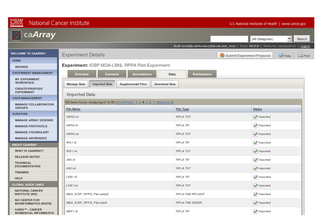
- **TCGA genomic characterizations**
- **GlaxoSmithKline cancer cell line panel**
- **Rembrandt**
- **TARGET**

Open Development

caArray is being extended to support reverse phase protein lysate arrays through a collaboration with Lawrence Berkeley National Laboratory.

MAGE-TAB ↓ RPLA-TAB
The MAGE-TAB data exchange standard is **extended** to support:
- Annotation of antibodies and their validation status
- "Reverse" array model, where spots are samples and the entire array is hybridized with one antibody

caArray ↓ caRPLA
The caArray2 data management system is **extended** to support:
- Parsing, validation, and import of an RPLA-TAB dataset
- Data and annotation management for antibodies and RPLAs
- Matched samples between microarray and RPLA assays



Knowledge Center

caBIG™ Enterprise Support Network

caArray is a component of the Molecular Analysis Knowledge Center at Columbia University and The Broad Institute of MIT and Harvard

Columbia University Medical Center
BROAD INSTITUTE