

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

Bian X¹, Klemm J¹, Basu A¹, Hadfield J¹, Srinivasa R², Parnell T², Miller S², Mason W², Kokotov D², Duncan M², Duvall P³, Gurses L³, Boal T⁴, Misquitta L⁴, Swan D⁵, Wysong R⁵, Klink A⁵, Johnson A⁵, Fontenay G⁶, Liu J⁷, Colbert M⁷, Komatsoulis G¹ ¹NCI Center for Biomedical Informatics and Information Technology, ²5AM Solutions, Inc., ³Stelligent, Inc., ⁴NARTec, Inc., ⁵TerpSys, Inc., ⁶Lawrence Berkeley National Laboratory, ⁷Science Applications International Corporation (SAIC) National Laboratory



Overview

caArray and MAGE-TAB

Integration with Data Analysis Tools

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

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

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

of the data submitted ensures data in high quality and abide by community standard for easy data sharing and exchanging



Home Page

Jan 21 10:11:53 EST 2009 | Node: NCICB | Welcome, caarr

array.nci.nih.gov 🔒

v study of cancer dene expression analysis using oligonucleotic

preliminary models. The performance of such confirmatory studies and the transition to clinical practice requi that microarray data from different laboratories are comparable and reproducible. We designed a study to as

metrix Human Genome U133A arrays. High within-laboratory and between-laboratory correlations or erved on the purified RNA samples, the cell lines, and the frozen tumor tissues. Intraclass correlation

boratories was only slightly stronger than between laboratories, and the intraclass correlation t akest for genes expressed at low levels and showing small variation. Finally, hierarchical cluster analy weakes for genes expressed a low levels and showing small variation. Initially, interactional cluster analysis revealed that the repeated samples clustered together regardless of the laboratory in which the experiments were done. The findings indicate that under properly controlled conditions it is feasible to perform complete tumor microarray analysis, from tissue processing to hybridization and scanning, at multiple independent laboratories for a single study. "NOTE: Migrated from caArray 1.x, identifier='gov.nih.nci.ncib.caarray:Experiment:1015897589951061:1'

Fissue Sites: Kidney, Lung, Stomach, Uterus, Liver, Lymphoid tissue, Ovary, Skin, Adrenal Gland, Lymph_Node

Experiment Overview Page

Disease States: Recurrent Renal Cell Carcinoma, Squamous Cell Carcinoma, Conventional_Clear_Cell_Renal_Cell_Carcinoma, Gastrointestinal_Stromal_Tumor, Lung_Adenocarcinoma, Leiomyoma, Non neoplastic liver with cirrosis, Stomach Adenocarcinoma, Large Cell Lymphoma, Ovarian Adenocarcinoma, Melanoma, Malignant G1 Stromal Tumor, Adrenal Cortical Adenoma, Metastatic Renal Cell Carcinoma, Malignant Melanoma

Material Types: synthetic_RNA, cell, nuclear_RNA, total_RNA, organism_part

comparability of data from four laboratories that will conduct a larger microarray profiling confirmation projei g adenocarcinomas. To test the feasibility of combining data across laboratories, frozen tumor tissues, cell li lets, and purified RNA samples were analyzed at each of the four laboratories. Samples of each type and

https://array.nci.nih.gov/caarray/project/details.action?project.id=

Overview Contacts Annotations Data Publications

Status: Draft Experiment Identifier: dobbi-0010

> Experiment URL: https://array.nci.r Assay Type*: Gene Expressio Provider*: Affymetrix Array Designs: HG-U133A

> > Organism*: Homo sapiens (nob

(periment: Interlabor

•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

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 Experimental Factor Type DISEASESTATE 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

	1			1					
#The first line be	low has be	en filled in w	ith example	e terms, where possible					
Source Name	Material Ty	Term Sourc	Characteri	Term Source REF	Characteristics[CellType]	Term Sour	Characteristics[Organism]	Characteri:	Term Sou
BT474 - Replicati	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa
BT474 - Replicati	Cell	MO	Breast	NCI Thesaurus	Epithelium	NCI Thesa	Homo sapiens	Carcinoma	NCI Thesa
BT474 - Replicati	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



•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

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

caArray in the Community

•Designed to encourage the rapid

•Current library includes over 100

•Gene Expression Analysis

Data Conversion Modules

workflows through its pipeline engine

•Supports sharing of analysis

integration of new techniques

analysis modules

•SNP Analysis



	Overview Contac	ts /	Annotations	Dat	a P	ublications	
lana	ge Data Imported Data	Suppl	emental Files	Downlo	ad Data		
)ov	vnload Data						
Filte	er By File Type: (All)		Fi	lter By S	itatus: (AII)	*	
•	<u>File Name</u>		File Type	<u>Ext.</u>	<u>Compressed</u> <u>Size</u>	<u>Uncompressed</u> <u>Size</u>	Download Queue <u>(Show Files</u>)
•	experiment- id-1015897559579654_curat	ed.idf	MAGE_TAB_IDF	.idf	2 KB	4 KB	O Files, Job Size: O Bytes
•	experiment- id-1015897559579654_curat	ed.sdrf	MAGE_TAB_SDRF	.sdrf	1 KB	18 KB	
Ð	am18_924_hu68_110200_ab	.CEL	AFFYMETRIX_CEL	.CEL	2.2 MB	7.1 MB	
)	cd34_6_hu68_160200_ab.CEL		AFFYMETRIX_CEL	.CEL	2.1 MB	6.9 MB	
•	aml_830_hu68_210400_ab.(CEL	AFFYMETRIX_CEL	.CEL	2.3 MB	7.1 MB	
Ð	am18_#5_hu68_110899_ab.0	CEL	AFFYMETRIX_CEL	.CEL	2.2 MB	7.1 MB	
Ð	od34_7_hu68_160200_ab.CB	ΞL	AFFYMETRIX_CEL	.CEL	2.2 MB	7 MB	
Э	aml_883_hu68_210400_ab.(CEL	AFFYMETRIX_CEL	.CEL	2.3 MB	7.1 MB	
•	aml_1028_hu68_210400_ab	.CEL	AFFYMETRIX_CEL	.CEL	2.3 MB	7.1 MB	
•	am18_294_hu68_110200_ab	.CEL	AFFYMETRIX_CEL	.CEL	2.2 MB	7 MB	
•	am18_#6_hu68_290799_ab.(CEL	AFFYMETRIX_CEL	.CEL	2.3 MB	7.2 MB	
•	aml_645_hu68_210400_ab.(CEL	AFFYMETRIX_CEL	.CEL	2.3 MB	7.2 MB	
•	am18_221_hu68_110200_ab	.CEL	AFFYMETRIX_CEL	.CEL	2.2 MB	7.1 MB	
Ð	aml8_c3_hu68_300799_ab.0	EL	AFFYMETRIX_CEL	.CEL	2.2 MB	7.1 MB	
	🔞 Clear Downloa	ad Queue	[📤 Launch D	ownloa	doL be	📤 Export Consoli	idated MAGE-TAB

	Contacts	Annotation	is	Data Pu	blications						
Experimental Design	Experimenta	l Factors	Sources	Samples Ext	acts Labeled E	dracts	Hybridizations				
Labeled Extracts								🙆 Ad	d a ne	w Label	led Extract
152 items found, displayi	ng 1 to 15. (Fire	st/Prev] 1, <u>2, 3</u>	<u>4, 5, 6, j</u>	7, 8[Next/Last]							
Labeled Extract name	Description	Material Type	Relate	d Extract	Hybridizations			Edit	Сору	Delete	Download
D001 normal prostate cRN	A	synthetic_RN/	A <u>D001</u>	normal prostate cDNA	PD001_u95b, PD001	u95c.,	PD001_u95av2			ü	*
D002 normal prostate cRN	A	synthetic_RN/	A <u>D002</u>	normal prostate cDNA	PD002_u95av2_PD0	102 u95c	: , <u>PD002_u95b</u>	2	5	i	
D003 normal prostate cRN	A	synthetic_RN/	A <u>D003</u>	normal prostate cDNA	PD003_u95c_, PD003 PD003_u95b	095Av	2.scan.2nd	2	•	ü	*
D004 normal prostate cRN	A	synthetic_RN/	Label	ed Extracts > D001_no	mal_prostate_cRNA		- 00004005-		-	etta	л ц .
D005 normal prostate cRN	A	synthetic_RN/	A Required	Required fields are marked with 'asterisks' .							
D006 normal prostate cRN	A	synthetic_RN/	4	Description: Extracts': D001_normal_prostate_cDNA Material Type: synthetic_RNA (MO)							
D007 normal prostate cRN	A	synthetic_RN/	4								
D008 normal prostate cRN	A	synthetic_RN/	4		Label: biotin	extraction					
D009 normal prostate cRN	A	synthetic_RN/	4			(Z Edit				
D010 normal prostate cRN	A	synthetic_RN/	4 Down	load Data							
D011 normal prostate cRN	A	synthetic_RN/	A Filter B	y File Type: (All)	¥		Filter By Status: (All)	*			
	A	synthetic_RN/		<u>File Name</u>	File Type	<u>Ext.</u>	Compressed Size	Incompresse	d Size	De	ownload Queue (She
D012 normal prostate cRN		eurthetic PN	O	PD001_u95av2.CEL	AFFYMETRIX_CEL	.CEL	2.8 MB 9	0.4 MB			Files, Job Size: O By
D012 normal prostate cRN	A	synuneuc_Riv		PD001_u95b.CEL	AFFYMETRIX_CEL	.CEL	2.7 MB	0.3 MB			
D012 normal prostate cRN D013 normal prostate cRN	<u>A</u>				AFFYMETRIX_CEL	.CEL	2.0 MB				
D012 normal prostate CRN D013 normal prostate CRN D014 normal prostate CRN	A	synthetic_RN/		PD001_0956.CEL	AFEYMETRIX CHR	CHP	T.0 (0D	0.01010			
D012 normal prostate CRM D013 normal prostate CRM D014 normal prostate CRM D015 normal prostate CRM	A A A	synthetic_RN		PD001_095av2mas5.CHP	AFFYMETRIX_CHP	.CHP	4.7 MB	10.9 MB			
D012 normal prostate CRN D013 normal prostate CRN D014 normal prostate CRN D015 normal prostate CRN	A A A	synthetic_RN/ synthetic_RN/		PD001_u95av2mas5.CHP PD001_u95bmas5.CHP PD001_u95bmas5.CHP PD001_u95cmas5.CHP	AFFYMETRIX_CHP AFFYMETRIX_CHP AFFYMETRIX_CHP	.CHP .CHP .CHP	4.7 MB 1	10.9 MB			

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

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			

Open Development

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

> 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





👫 National Ca	ncer Institute		U.S. National Institutes of Health www.cancer.gov
📝 caArray			(All Categories) 🔽 Search
		Build trunk(\${buildNumber}) \${bu	uild.date_time} Node: NCICB Welcome, caarrayadmin Logou
VELCOME TO CAARRAY	Experiment Details		Submit Experiment Proposal @ Help + Prin
HOME			
BROWSE	Experiment: ICBP MDA-LBNL RPPA Pilot Experiment		
EXPERIMENT MANAGEMENT	Overview Contacts Annotations	ata Publications	
MY EXPERIMENT WORKSPACE			
CREATE/PROPOSE EXPERIMENT	Impred Data Supplemental Pres Download L	Ge Bal	
ISER MANAGEMENT			
MANAGE COLLABORATION	165 items found, displaying 61 to 80. [First/Prev] 1, 2, 3, 4, 5, 6, 7, 8[Next/Lat		
URATION	File Name	File Type	Status
MANAGE ARRAY DESIGNS	HER2.txt	RPLA TXT	M Imported
MANAGE PROTOCOLS	HER3.tif	RPLA TIF	Marked Imported
MANAGE VOCABULARY	HER3.txt	RPLA TXT	M Imported
MANAGE ANTIBODIES	ID:14	DDI A TIE	
BOUT CAARRAY	ino i.u	AFLA IIF	M Imponed
WHAT IS CAARRAY?	IRS1.txt	RPLA TXT	M Imported
RELEASE NOTES	JNK.tif	RPLA TIF	M Imported
DOCUMENTATION	JNK.txt	RPLA TXT	Minported
TRAINING		DDLA TE	
HELP	LNB1.0	HPLA IIF	Imported
LOBAL QUICK LINKS	LKB1.txt	RPLA TXT	Imported
INSTITUTE (NCI)	MDA_ICBP_RPPA_Pilot.mplaidf	RPLA-TAB RPLAIDF	Imported
NCI CENTER FOR BIOINFORMATICS (NCICB)	MDA_ICBP_RPPA_Pilot.sradf	RPLA-TAB SRADF	Imported
CABIG TH - CANCER BIOMEDICAL INFORMATICS	MEK1.tif	RPLA TIF	Imported

Knowledge Center

caBIG[®] Enterprise Support Network

Molecular Analysis Knowledge Center

at Columbia University and The Broad

Institute of MIT and Harvard