

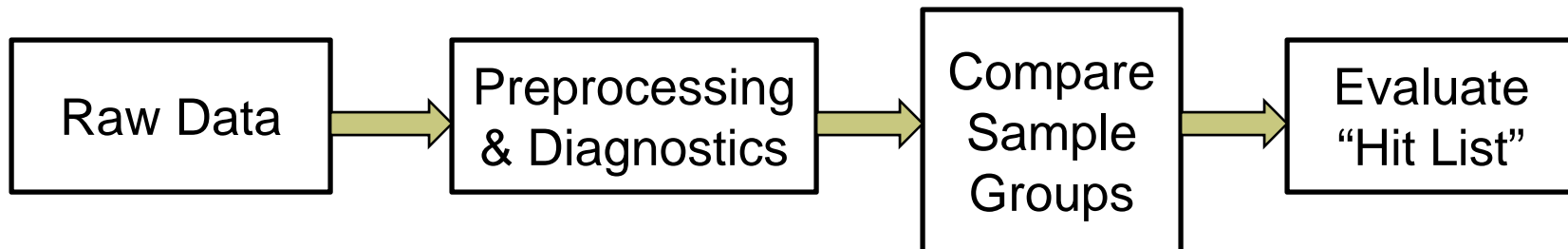


# Maine INBRE Bioinformatics Core Training Programs - Microarray Workshop I -

April 30, 2009

# Goal of Microarray Workshop I

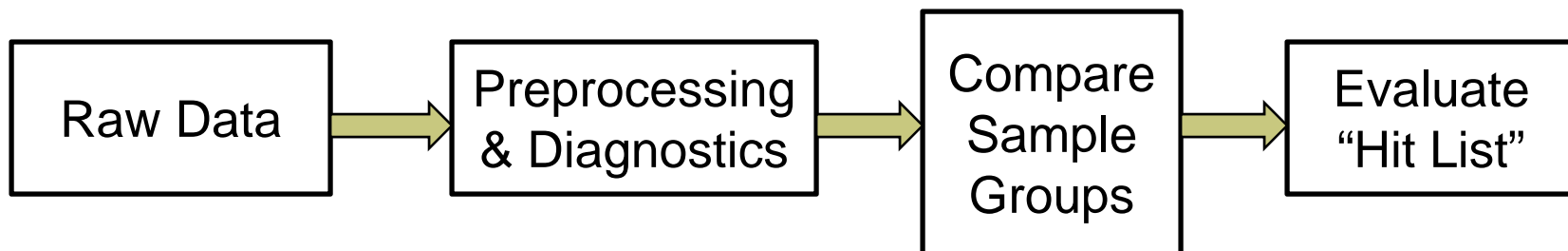
- Provide training for the analysis of microarray gene expression data
  - Use statistical computing environment, R
    - Bioconductor
    - R/maanova (Dr. Gary Churchill)
  - Two specific array platforms:
    - Affymetrix Mouse 430v2 GeneChip
    - Agilent Zebrafish 44k array



# Schedule

## Program

	Topic	Instructor
9:00-10:00	Introduction	Ben King
10:00-12:00	Normalization	Tim Stearns
12:00 -1:00	Lunch	
1:00-3:00	Analysis: Gene selection	Weidong Zhang
3:00-5:00	Analysis of Gene lists: functional annotation and network prediction	Ben King



# Course Materials

- Course web site
  - [http://www.maineidea.net/Core/bioinformatics\\_workshop.html](http://www.maineidea.net/Core/bioinformatics_workshop.html)
- R installation instructions
  - Bioconductor and R/maanova
- Files for Affymetrix Mouse 430v2 Experiment (published)
  - Data files
  - R script
- Files for Agilent Zebrafish 44k Experiment (unpublished)
  - Data files (please don't distribute)
  - R script

# Outline of Introduction

- Microarray platforms
- Experimental design
- Overview of analysis workflow
- Raw data formats
- Major data repositories
- Shopping for data

# Microarray Platforms

- Two-color
  - Two samples simultaneously assayed per array using different dyes
    - Green dye – Cyanine 3 (Cy3) excited maximally at 550nm
    - Red dye – Cyanine 5 (Cy5) excited maximally at 649nm
  - e.g., Agilent Zebrafish 44k array
- One-color
  - One sample assayed per array
  - e.g., Affymetrix Mouse 430v2 array

# Example One-Color Platform Experiment

## *Pparg*<sup>ldi</sup> Mutant Adipocytes



## Wild-type Control Adipocytes

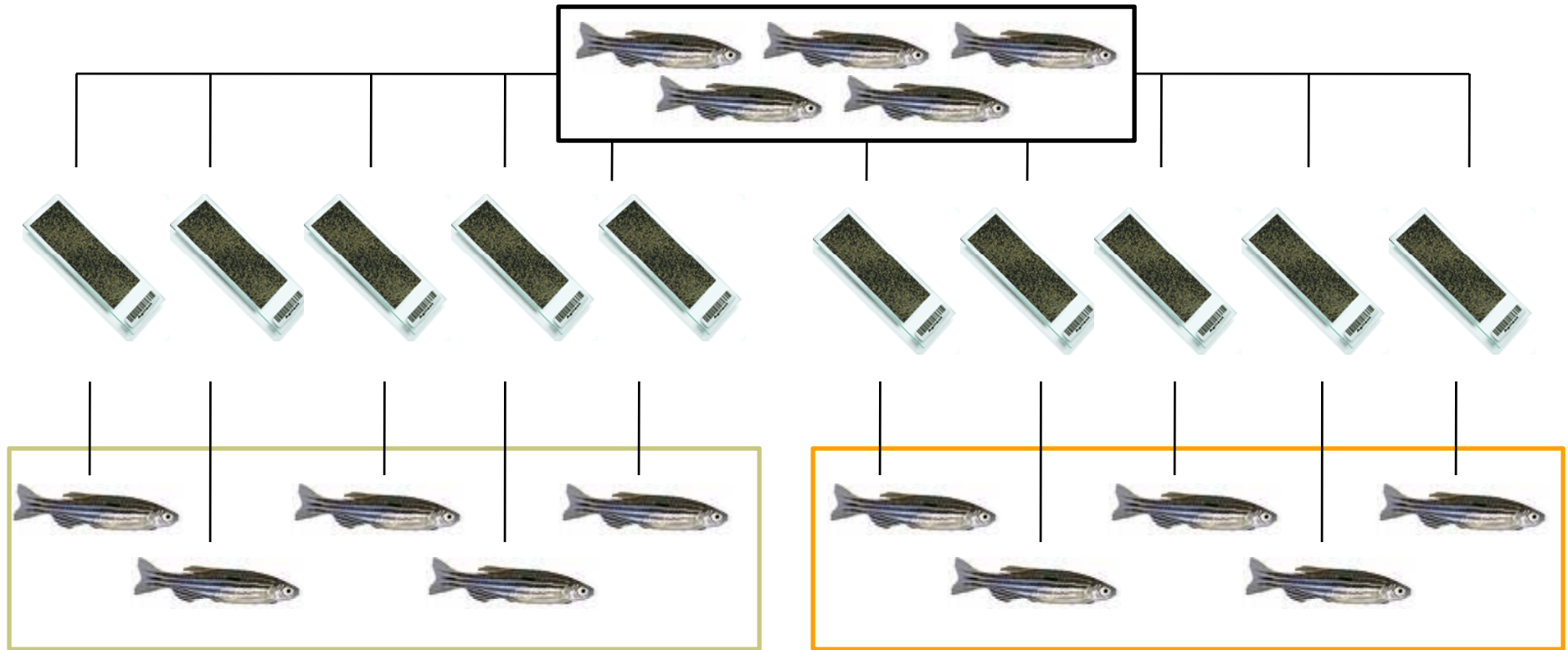


RNA samples assayed using Affymetrix 430 v2 GeneChips

Kim *et al.* *PNAS* (2007) 104(42): 16627-32.

# Example Two-Color Platform Experiment

RNA Reference Pool (Cy3)



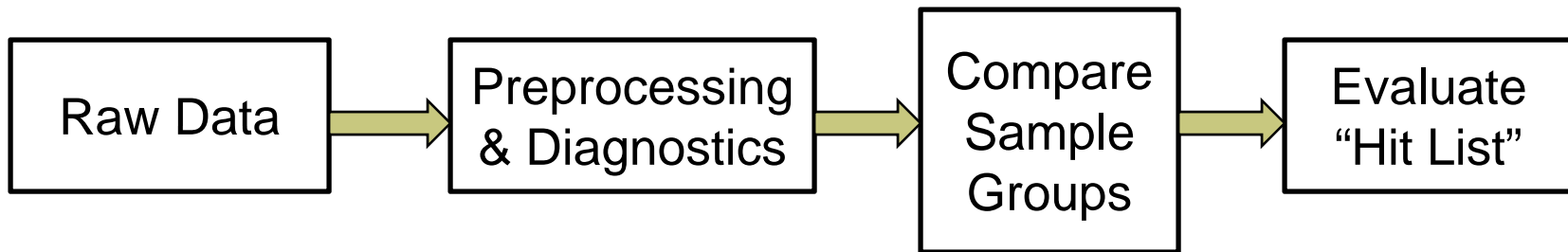
Arsenic Treated  
(Test Group, Cy5)

Water Treated  
(Control Group, Cy5)

# Experimental Design

- Attempt to control for as many factors as practically possible:
  - e.g., Age and sex match
- Biological replication is very important
  - Use more replicates when response is variable or want more sensitive detection
    - e.g., may use 3 replicates in inbred mice
    - e.g., may use 10 replicates to study disease progression
  - Balanced number of replicates
- Technical replication often used with two-color platforms
  - Dye-flip

# Microarray Data Analysis Workflow



- Raw Data
  - Get intensities for each probe or probe set
  - Document design
- Data Preprocessing and Diagnostics
- Compare Sample Groups (ANOVA)
- Evaluate "Hit List"

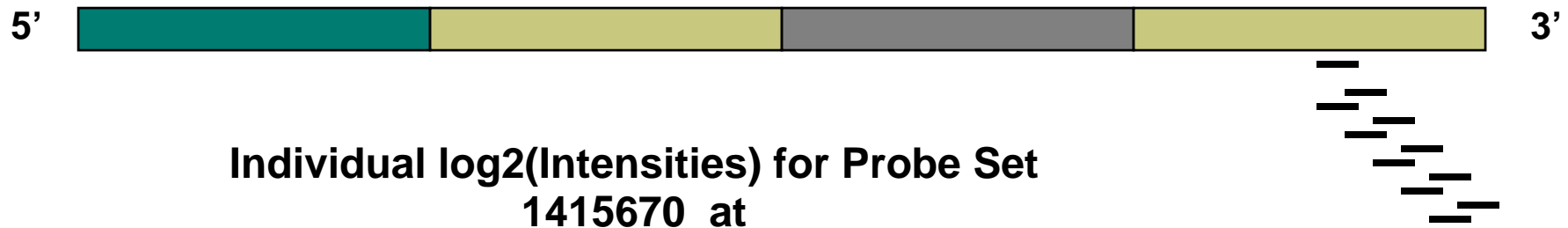
# Get Intensities From Raw Data

- Use raw intensities for diagnostics
  - Look for spatial bias or other anomalies in data
  - Explore normalization methods
  - Examine variance within and between sample groups
- Use processed intensities to compare sample groups

# Affymetrix Raw Data Files (Mouse 430v2 Gene Chip)

- Raw Data
  - Intensities stored in .CEL files (binary format)
  - Require a Chip Description File (.CDF)
    - Location of all 25-mer probes on array
    - Describes how groups of 11 probes map to a probe set (gene)
- Summarized Data
  - “Average” intensity for each probe set
- Annotation Data
  - Annotation for each probe set available on Affymetrix web site (NetAffx)

# Summarization of Affy Probe Intensities



Two summarization methods:

- Robust Multi-Chip Averaging (RMA) – Use Perfect Match (PM) probes
- MAS5 (.CHP file)
  - Uses intensities from Perfect Match (PM) and Mismatch (MM) Probes

## Agilent Raw Data Files (Agilent Zebrafish v2 44k array)

- Agilent feature extraction file (plain text)
  - Values computed from scanner image files
    - Cy3 and Cy5 foreground intensities in separate columns
  - Contains annotation
- Other two-color platforms often use GenePix software for feature extraction
  - Generates intensities from scanner image files
  - Requires a .GAL file that describes probe location on array (block, row, column)
  - End result is a plain text file

# Major Microarray Data Repositories

- Most journals require that microarray data be deposited prior to publication
- Gene Expression Omnibus (GEO) at NCBI
  - Platform record (GPLXXXXX)
    - Can be multiple for same array
  - Sample record (GSMXXXXX) – Describes samples on an array
  - Series record (GSEXXXXX) – Describes experiment
  - GEO Data Sets and GEO Profiles
- ArrayExpress at EBI
  - Independent repository from GEO - may have some unique data
  - MAGE-TAB files to describe experimental design

NCBI  
Gene Expression Omnibus

HOME SEARCH SITE MAP Handout NAR 2006 Paper NAR 2002 Paper FAQ MIAME Email GEO

NCBI > GEO Not logged in | Login

**Gene Expression Omnibus:** a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

**GEO navigation**

**QUERY**

- DataSets:
- Gene profiles:
- GEO accession:
- GEO BLAST

**BROWSE**

- DataSets
  - Platforms
- GEO accessions
  - Samples
  - Series

**SUBMIT**

- Direct deposit / update
- Web deposit / update
- Create new account

**Public data**

GPL Platforms	5822
GSM Samples	302973
GSE Series	11937
Total	320732

**Site contents**

**Documentation**

- Overview | FAQ | Find
- Submission guide
- Linking & citing
- Journal citations
- Programmatic access
- DataSet clusters
- GEO announce list
- Data disclaimer
- GEO staff

**Query & Browse**

- Repository browser
- Submitter contacts
- SAGEmap
- FTP site
- GEO Profiles
- GEO DataSets

**Deposit & Update**

- Direct deposit
- Web deposit

Search by keyword  
(e.g., author)

Browse platforms and  
view experiments

mouse Il1b - GEO Profiles Results - Mozilla Firefox

http://www.ncbi.nlm.nih.gov/sites/entrez

NCBI GEO PROFILES Gene Expression Omnibus

Search GEO Profiles for mouse Il1b

Display Summary Show 20 Sort by Subgroup effect

All: 45924

Items 1 - 20 of 45924 Page 1 of 2297

1: GDS2686 record | GPL1261 1449399\_a\_at [Mus musculus] 4 samples

Annotation: Il1b: interleukin 1 beta RP23-384K11.2, IL-1beta, IL-1b

Reporter: BC011437

Experiment: MyD88 deficient macrophage response to zymosan, gene expression array-based, transformed count

2: GDS164 record | GPL207 F4k [Mus musculus] 6 samples

Annotation: Il1b: interleukin 1 beta RP23-384K11.2, IL-1beta, IL-1b

Reporter: M15131

Experiment: Shock and adaptive response to injury, gene expression array-based, count

3: GDS859 record | GPL81 103486\_at [Mus musculus] 4 samples

Annotation: Il1b: interleukin 1 beta RP23-384K11.2, IL-1beta, IL-1b

Reporter: M15131

GEO DataSet Browser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2225

Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers

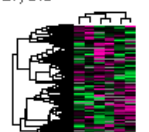
NCBI DATASET BROWSER CURATED GEO Gene Expression Omnibus

Search for GDS2225[ACCN] Search Clear Show All Advanced Search

**DataSet Record GDS2225:** Expression Profiles Data Analysis Tools Sample Subsets

<b>Title:</b>	Mechanical strain effect on fetal lung type II epithelial cells		
<b>Summary:</b>	Analysis of embryonic day 19 fetal lung type II epithelial cells exposed to mechanical strain for 16 hours. Mechanical forces are essential for normal fetal lung development. Results provide insight into the mechanisms regulating lung development in response to mechanical forces.		
<b>Organism:</b>	<i>Rattus norvegicus</i>		
<b>Platform:</b>	GPL341: [RAE230A] Affymetrix Rat Expression 230A Array		
<b>Citation:</b>	Wang Y, Maciejewski BS, Weissmann G, Silbert O et al. DNA microarray reveals novel genes induced by mechanical forces in fetal lung type II epithelial cells. <i>Pediatr Res</i> 2006 Aug;60(2):118-24. PMID: 16864689		
<b>Reference Series:</b>	GSE3541	<b>Sample count:</b>	6
<b>Value type:</b>	count	<b>Series published:</b>	2006/04/29

**Cluster Analysis**



**Download**

- DataSet SOFT file
- Series family SOFT file
- Series family MINIML file
- Annotation SOFT file

**Data Analysis Tools**

Find genes ?

Compare 2 sets of samples

Cluster heatmaps

Experiment design and value distribution

Find gene name or symbol:  Go

Find genes that are up/down for this condition(s):  stress Go

ArrayExpress Home - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ebi.ac.uk/microarray-as/ae/

Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

# ARRAYEXPRESS

ArrayExpress is a public archive for **transcriptomics data**, which is aimed at storing **MIAME-** and **MINSEQE-** compliant data in accordance with MGED recommendations. The ArrayExpress Warehouse stores gene-indexed **expression profiles** from a curated subset of experiments in the archive.

[» More Info](#)

## Experiments Archive

8012 experiments, 236608 assays

Experiment, citation, sample and factor annotations

[Browse experiments](#)  
[Advanced query interface](#)

Query

[Submitter/reviewer login](#)

[ArrayExpress Query Help](#)

## Atlas of Gene Expression

946 experiments, 26530 assays, 4911 conditions

Genes

Conditions

up/down in

Any species

Query

[ArrayExpress Atlas Home](#)

[Query ArrayExpress Warehouse](#)

## News

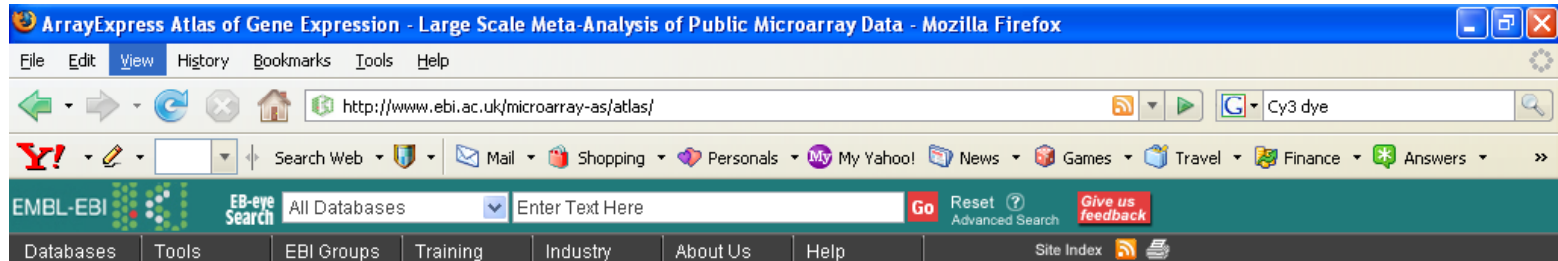
### 02 Mar 2009 - ArrayExpress Atlas Student Masters Project

An MSc Masters project placement is available within the Microarray Group. It will involve developing an advanced ontology querying interface to ArrayExpress Atlas. For further details, please check [this page](#).

## Links

- ◆ [ArrayExpress User Survey](#)
- ◆ [Help](#) | [Training](#) | [FAQ](#) | [Citing](#)
- ◆ [Submit data](#) to ArrayExpress
- ◆ [Programmatic Access](#) | [FTP Access](#)
- ◆ [Software Downloads](#) and [Statistics](#)
- ◆ [EFO](#) | [Bioconductor package](#) | [Quality metrics](#)
- ◆ [ArrayExpress Scientific Advisory Board](#)
- ◆ [Microarray Informatics Group](#)

[Terms of Use](#) | [EBI Funding](#) | [Contact EBI](#) | © European Bioinformatics Institute 2009. EBI is an Outstation of the [European Molecular Biology Laboratory](#).



# ATLAS

[about the project](#) | [faq](#) | [feedback](#) | [blog](#) | [web services api](#) | [help](#)

Genes:  e.g. ASPM, "p53 binding"
 Organism:  e.g. liver, cancer, diabetes
 Conditions:  e.g. liver, cancer, diabetes

[show help](#)  
[advanced search](#)

**Atlas Data Release 9.3:**

new experiments	80
total experiments	946
assays	26530
conditions	4911

**ArrayExpress Atlas of Gene Expression**

ArrayExpress Atlas is a semantically enriched database of meta-analysis based summary statistics over a curated subset of ArrayExpress Archive, servicing queries for condition-specific gene expression patterns as well as broader exploratory searches for biologically interesting genes/samples.

For news and updates, subscribe to the [atlas mailing list](#).

# Shopping For Data

- Best experiments will feature biological replication
- Want raw data (.CEL files or text files with intensities)
- Need information about experimental design
  - How each sample maps to a raw data file

Platforms (1) [GPL1261](#) [Mouse430\_2] Affymetrix Mouse Genome 430 2.0 Array

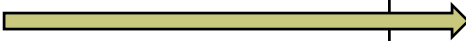
Samples (6) [GSM231015](#) Pparg-Idi\_MUT1  
[GSM231016](#) Pparg-Idi\_MUT2  
[GSM231018](#) Pparg-Idi\_MUT3  
[GSM231019](#) Pparg-Idi\_WT1  
[GSM231020](#) Pparg-Idi\_WT2  
[GSM231021](#) Pparg-Idi\_WT3

This SubSeries is part of SuperSeries:  
[GSE9132](#) Identifying gene expression changes in adipose tissue of lipodystrophic mice

Download family	Format
<a href="#">SOFT formatted family file(s)</a>	SOFT <a href="#">?</a>
<a href="#">MINiML formatted family file(s)</a>	MINiML <a href="#">?</a>
<a href="#">Series Matrix File(s)</a>	TXT <a href="#">?</a>

Supplementary file	Size	Download	File type/resource
<a href="#">GSE9131_RAW.tar</a>	38.3 Mb	<a href="#">(ftp)</a> <a href="#">(http)</a>	TAR (of CEL)

Raw data provided as supplementary file  
Processed data included within Sample table



	A	B	C	D
1	Array	Strain	Dye	Sample
2	GSM231015	MUT	1	1
3	GSM231016	MUT	1	2
4	GSM231018	MUT	1	3
5	GSM231019	WT	1	4
6	GSM231020	WT	1	5
7	GSM231021	WT	1	6
8				
9				



R/maanova design file