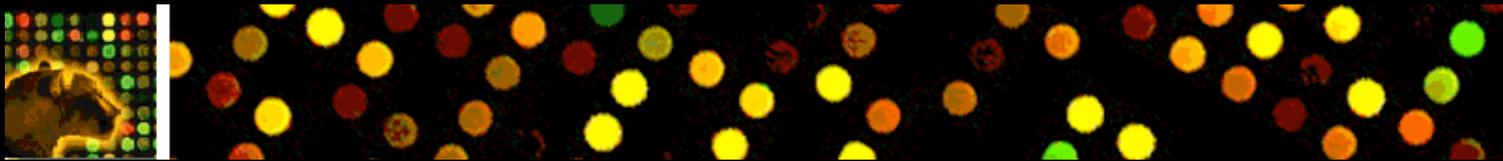


# Welcome to PUMAdb



## Princeton University MicroArray database

October 17, 2008

John Matese



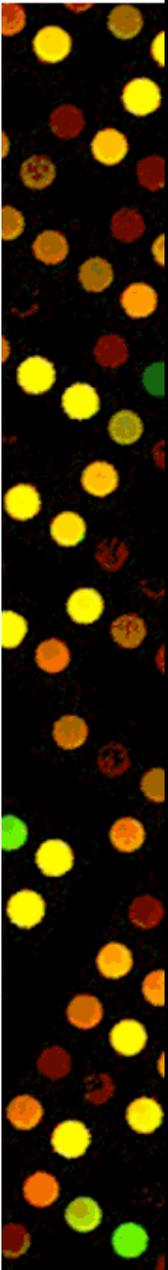
# User Help: Tutorials and Workshops

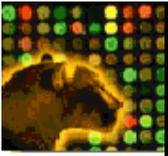
- Help & FAQ
  - <http://puma.princeton.edu/help/>
  - <http://puma.princeton.edu/help/FAQ.shtml>
- Tutorials
  - [http://puma.princeton.edu/help/tutorials\\_subpage.shtml](http://puma.princeton.edu/help/tutorials_subpage.shtml)
  - Ideas? Email [array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)
- Hybridization & Scanning Individual Instruction
  - Email [dstorton@molbio.princeton.edu](mailto:dstorton@molbio.princeton.edu)



# Welcome to the database: a tutorial

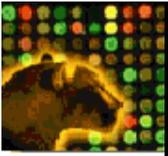
- What we'll talk about:
    - User Registration
    - Staging Data
    - Loading Prerequisites
    - Loading Data
    - Finding Your Data
    - Displaying Your Data
    - Data Retrieval and Analysis
    - Organizing Data
    - Submitting Plate Samples
  - What we will not discuss, or only brush the surface of...
    - Experimental Design
    - Experimental Protocol
    - Data Normalization
    - Data Quality Assessment
    - Data Analysis (clustering)
    - External User Tools (XCluster, TreeView, etc.)
- 
- Please fill out the sign-up sheet and survey form
  - Questions? email us at: [array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)





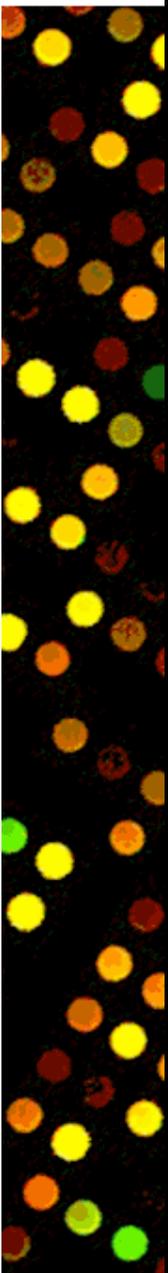
# Welcome to PUMAdb

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Data Retrieval and Analysis
- Organizing Data
- Submitting Plate Samples



# User Registration

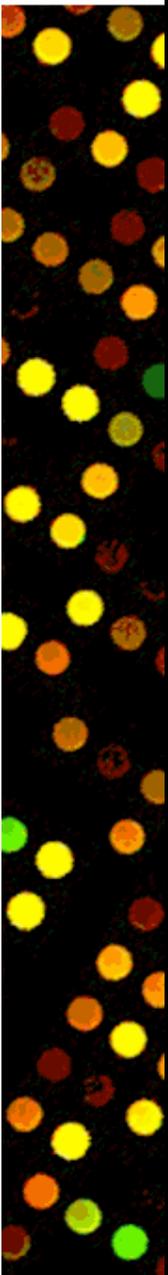
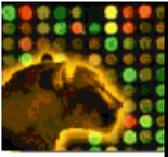
- PUMAdb is free
- Fill out the registration form
  - <http://puma.princeton.edu/cgi-bin/tools/display/registration.pl>
- Lab head (PI) should also register
- External collaborators can also be granted accounts (with different levels of access)
- Accounts are occasionally granted to external researchers without an existing on-site collaboration, provided support demands are reasonable.



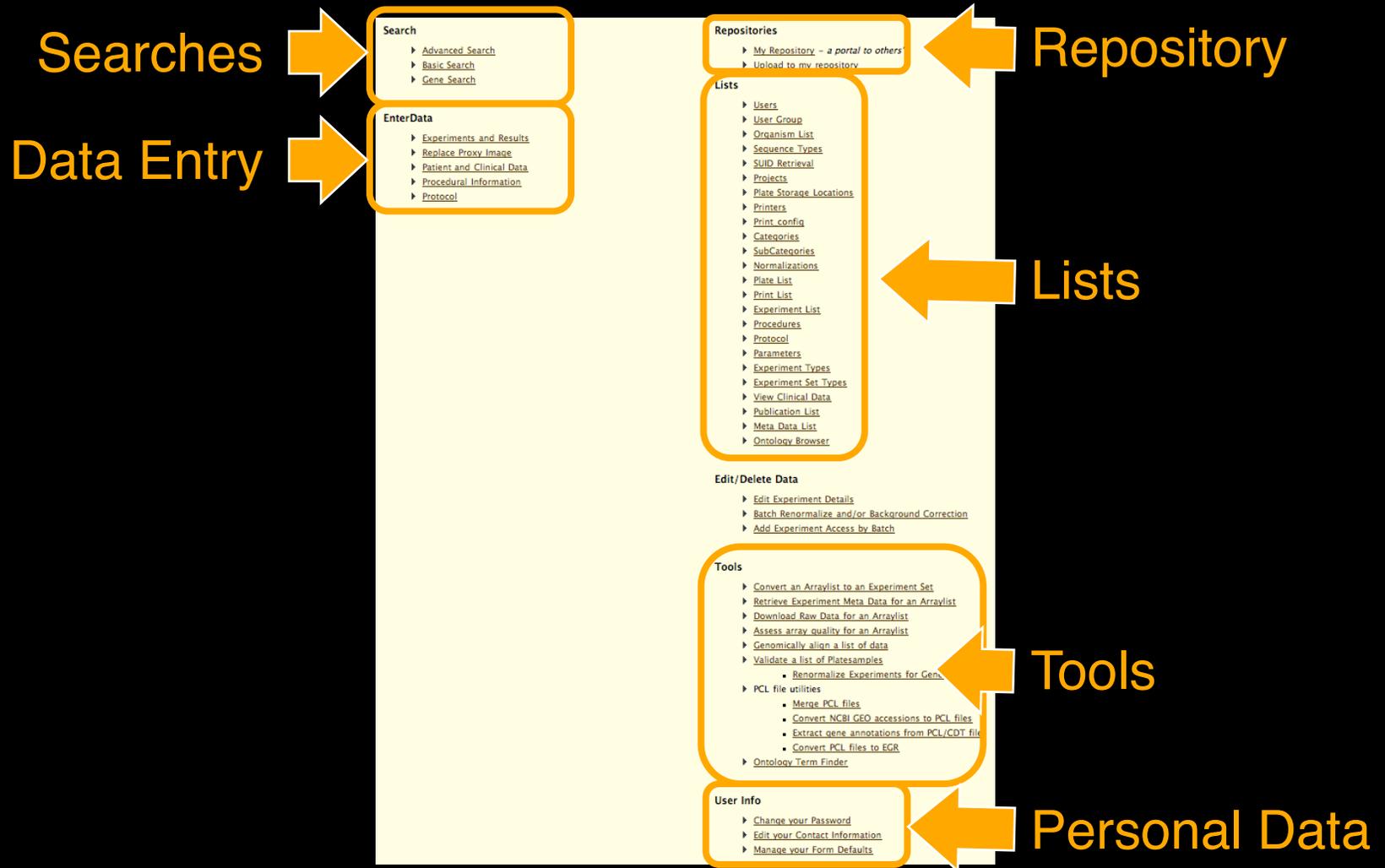


# Important Access Policies

- The experimenter for an assay retains all edit and delete privileges, solely
- All users within an access group (i.e. “lab”) can see each others data
- Upon publication, experiments are made public



# User Privileges: All Programs

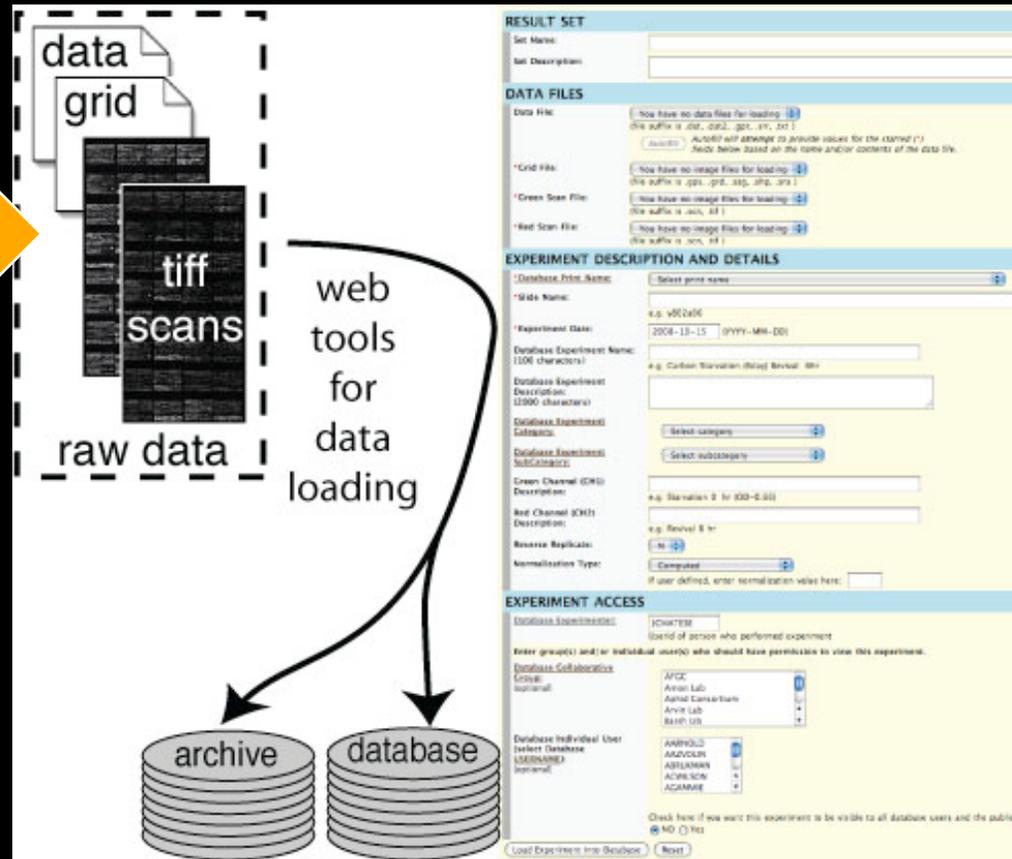


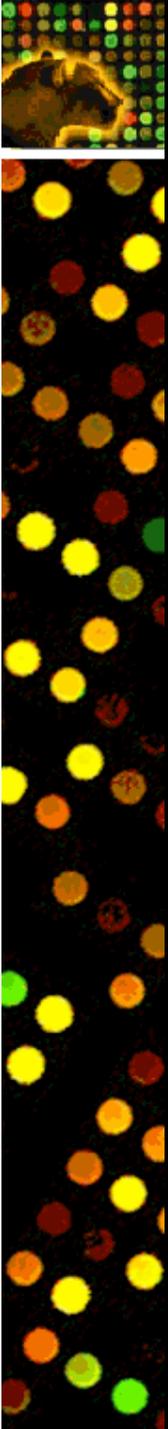


# Welcome to PUMAdb

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Data Retrieval and Analysis
- Organizing Data
- Submitting Plate Samples

# Submitting Data





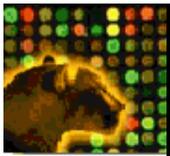
# Array Files vs. Loader

- Array Files, a SMB server
  - for core facility users, only
  - `smb://arrayfiles/arraydata`
  - Files within personal directory are visible by database automatically
- Loader, a SFTP server
  - for all unrestricted database users
  - With SFTP client, connect to `loader.princeton.edu` using `PUMAdb userid|password` and upload your data files to the 'incoming' directory



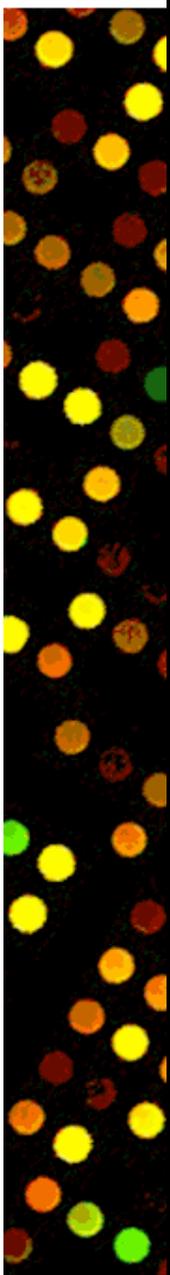
# Loader Account: Directories

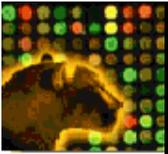
- **incoming**  
Stores all files prior to experiment loading. This is temporary storage - eventually, files will be deleted!
- **logs**  
Feedback files from the database are written to this directory (i.e. experiment loading logs)
- **arraylists**  
The database will look in this folder to retrieve any arraylists (list of arrays you have grouped together)
- **genelists**  
The database will look in this folder to retrieve any genelists (a list of genes with possible annotations)



# Welcome to PUMADB

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Data Retrieval and Analysis
- Organizing Data
- Submitting Plate Samples





# Loading Prerequisites

- Array design
  - Check the existing list of prints, and inform us if not found.
  - If you are using arrays from the core facility, this will be done for you
  - If you are creating your own prints (homemade contact-printed), please stay for the last 15 minutes of the tutorial...
- Experimental category and subcategory
  - Check the existing lists, and inform us if not found.
  - Make sure that your categories and subcategories are meaningful and not cryptic.

If new entries are required, email :

[array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)

# Loading Prerequisites : Lists

Search	List
<ul style="list-style-type: none"><li>▶ <a href="#">Advanced Search</a></li><li>▶ <a href="#">Basic Search</a></li><li>▶ <a href="#">Gene Search</a></li></ul>	<ul style="list-style-type: none"><li>▶ <a href="#">Users</a></li><li>▶ <a href="#">User Group</a></li><li>▶ <a href="#">Organism List</a></li><li>▶ <a href="#">Sequence Types</a></li><li>▶ <a href="#">SUID Retrieval</a></li><li>▶ <a href="#">Projects</a></li><li>▶ <a href="#">Plate Storage Locations</a></li><li>▶ <a href="#">Printers</a></li><li>▶ <a href="#">Print_config</a></li><li>▶ <a href="#">Categories</a></li><li>▶ <a href="#">SubCategories</a></li><li>▶ <a href="#">Normalizations</a></li><li>▶ <a href="#">Plate List</a></li><li>▶ <a href="#">Print List</a></li><li>▶ <a href="#">Experiment List</a></li></ul>
<b>EnterData</b> <ul style="list-style-type: none"><li>▶ <a href="#">Experiments and Results</a></li><li>▶ <a href="#">Replace Proxy Image</a></li><li>▶ <a href="#">Patient and Clinical Data</a></li><li>▶ <a href="#">Procedural Information</a></li><li>▶ <a href="#">Protocol</a></li></ul>	<ul style="list-style-type: none"><li>▶ <a href="#">Procedures</a></li><li>▶ <a href="#">Protocol</a></li><li>▶ <a href="#">Parameters</a></li><li>▶ <a href="#">Experiment Types</a></li><li>▶ <a href="#">Experiment Set Types</a></li><li>▶ <a href="#">View Clinical Data</a></li><li>▶ <a href="#">Repository List</a></li><li>▶ <a href="#">Publication List</a></li><li>▶ <a href="#">Meta Data List</a></li><li>▶ <a href="#">Ontology Browser</a></li></ul>

Existing categories & subcategories

Existing prints



# Welcome to PUMADB

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Data Retrieval and Analysis
- Organizing Data
- Submitting Plate Samples



# Annotation : Data Standards

- MGED - Micro Array Gene Expression Database Society
- “Minimal Information About a Microarray Experiment” (MIAME)
  - Experimental Design
  - Array Design
  - Biological Samples
  - Hybridizations
  - Measurements
  - Data Normalization and Transformation

*Nature Genetics (2001) 29, 365-371.*



# Annotation : MIAME Checklist

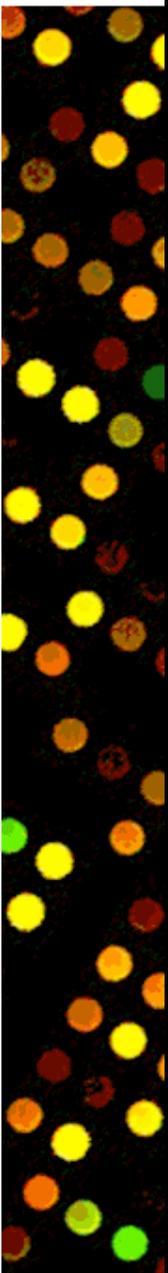
- In September 2002, MGED sent out a letter to journals and reviews requesting the microarray publications have this “minimal” information/annotation
- Many journals now have policies requiring published data to be well-annotated and deposited in a public repository (i.e. NCBI GEO).

<http://www.mged.org/Workgroups/MIAME/miame.html>



# Loading Data : Required Files

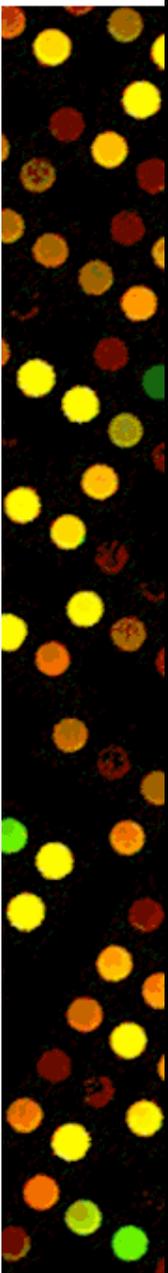
- In order to submit data, you need the following files staged:
  - For Affymetrix Data (dChip/GCOS/MAS5)
    - probeset\_data.txt, cell\_data.cel, experiment.exp, image.dat
  - For Agilent Data
    - data.txt, shape.shp, channel1.tif, channel2.tif
  - For GenePix Data
    - data.gpr, grid.gps, channel1.tif, channel2.tif
  - For Nimblegen Data
    - genes.txt|genes.calls, cell\_data.xys, features.ftr, image.tif
  - For ScanAlyze Data
    - data.dat, grid.sag, channel1.scn, channel2.scn





# Loading Data : Problems

- File names should only include allowed characters:
  - numbers, letters, dots, hyphens, underscores
  - **Spaces** and **slashes** are not allowed
- Images **must** be binary file (transfer issues)
- Only tif files may be compressed at the time of loading



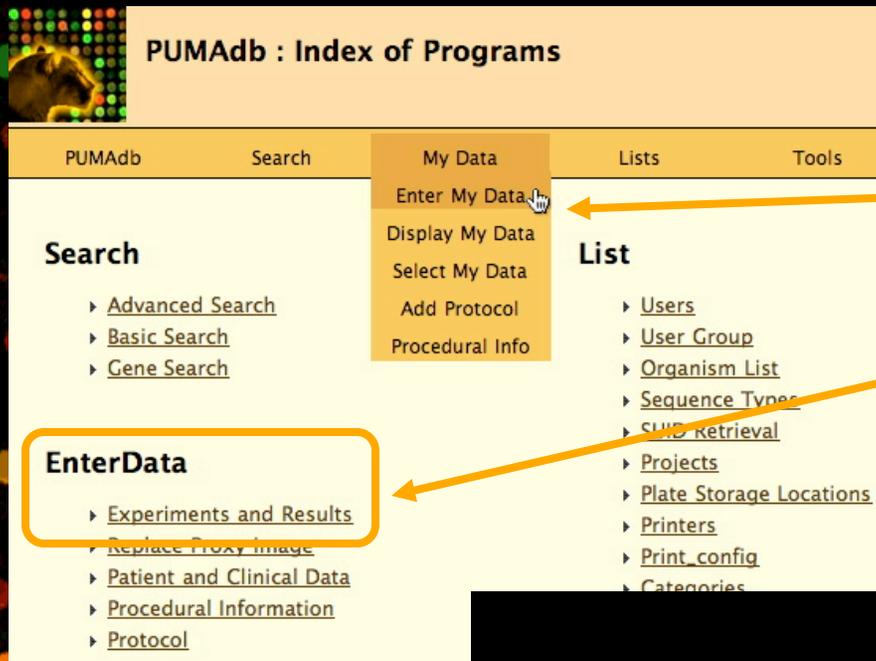


# Loading Data to the Database

- The *incoming* directory is emptied automatically every few weeks.
- Although we do archive your data, it does not serve as a raw datafile retrieval service, as of yet.

**Please Archive Your Data!**

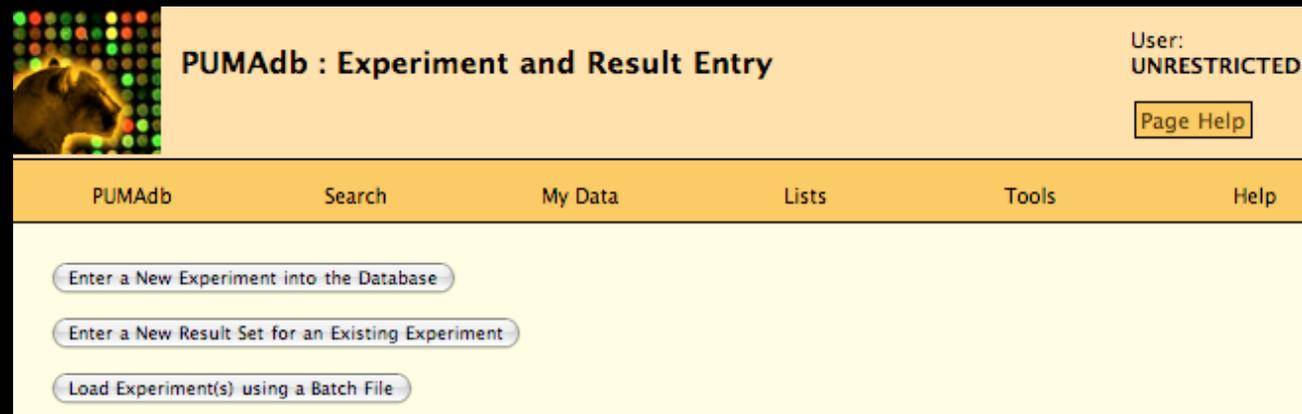
# Loading Data : Data Entry



The screenshot shows the PUMAdb website interface. The main navigation bar includes 'PUMAdb', 'Search', 'My Data', 'Lists', and 'Tools'. The 'My Data' menu is expanded, showing options: 'Enter My Data', 'Display My Data', 'Select My Data', 'Add Protocol', and 'Procedural Info'. The 'Enter My Data' option is highlighted with a mouse cursor. Below the navigation bar, there are two main sections: 'Search' and 'List'. The 'Search' section includes links for 'Advanced Search', 'Basic Search', and 'Gene Search'. The 'List' section includes links for 'Users', 'User Group', 'Organism List', 'Sequence Types', 'SUID Retrieval', 'Projects', 'Plate Storage Locations', 'Printers', 'Print config', and 'Categories'. A red box highlights the 'EnterData' section, which contains links for 'Experiments and Results', 'Replace Proxy Image', 'Patient and Clinical Data', 'Procedural Information', and 'Protocol'. Two red arrows point from the text on the right to the 'Enter My Data' option and the 'Experiments and Results' link.

- Choose your method
  - Within navigation menu...
  - Experiments and results link...

# Loading Data: Step 1



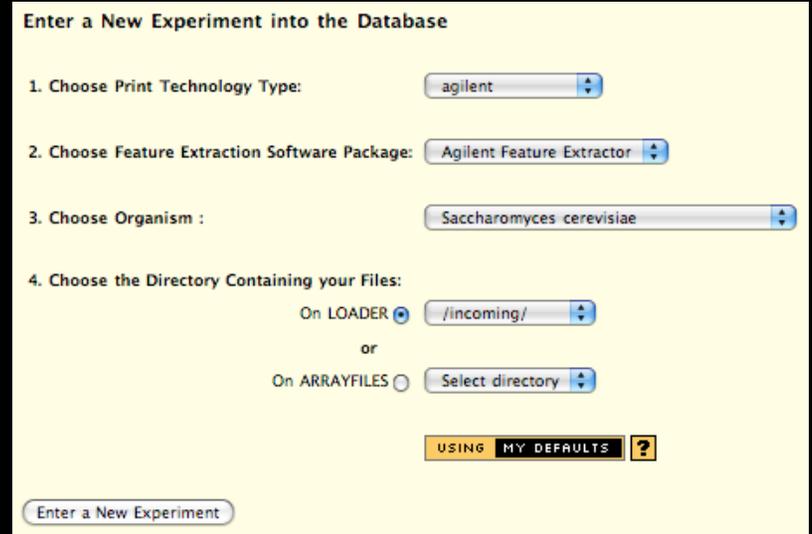
The screenshot shows the PUMAdb website interface. At the top left is a small image of a mouse head. The main header area contains the text "PUMAdb : Experiment and Result Entry" and "User: UNRESTRICTED" with a "Page Help" button. Below the header is a navigation menu with links for "PUMAdb", "Search", "My Data", "Lists", "Tools", and "Help". The main content area features three buttons: "Enter a New Experiment into the Database", "Enter a New Result Set for an Existing Experiment", and "Load Experiment(s) using a Batch File".

- Decide if you are entering a single experiment or a batch of experiments
- In specialized cases, add additional result sets for existing experiments



# Loading Data: Step 2

- Select the print technology (agilent, affymetrix, nimblegen, spotted)
- Select the feature extraction software package was used to generate your data
- Select the organism whose genes are arrayed
- Select the location of staged data



Enter a New Experiment into the Database

1. Choose Print Technology Type:

2. Choose Feature Extraction Software Package:

3. Choose Organism :

4. Choose the Directory Containing your Files:

On LOADER

or

On ARRAYFILES



# Loading Data: Result set

You are entering data for *Saccharomyces cerevisiae* into the database.  
All fields on this form are required except where indicated.

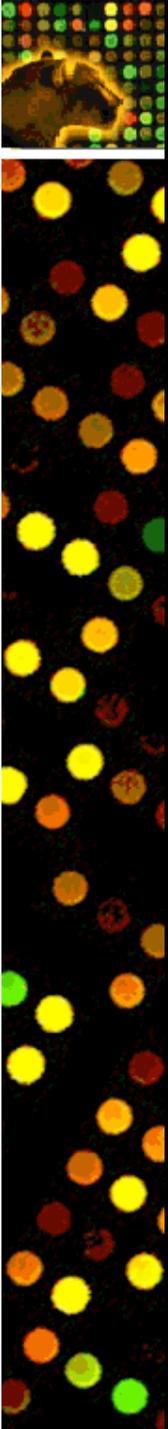
Only files in the directory you previously specified are displayed.

## RESULT SET

Set Name:

Set Description:

- For Affymetrix, Agilent, and Nimblegen results
- Provide a Result Set Name and Description
  - As for any single experiment there may be  $n$  result sets derived from raw data. You must create a name for each of these sets so that each result set may be identified and retrieved unambiguously from the database



# Loading Data : Data File Locations

## DATA FILES

Data File:  (file suffix is .dat, .dat2, .gpr, .srr, .txt )

*Autofill will **attempt** to provide values for the starred (\*) fields below based on the name and/or contents of the data file.*

\*Grid File:  (file suffix is .gps, .grd, .sag, .shp, .sra )

\*Green Scan File:  (file suffix is .scn, .tif )

\*Red Scan File:  (file suffix is .scn, .tif )

- Choose the data, grid, green scan and red scan files to be loaded from your staging directory
- Each pull down menu should be populated with the appropriate file
- Autofill button may help select the correct file, given a data file to start...

# Loading Data : Platform & Experiment Details

**EXPERIMENT DESCRIPTION AND DETAILS**

\* Database Print Name:

\* Slide Name:

\* Experiment Date:  (YYYY-MM-DD)

Database Experiment Name:   
(100 characters)

Database Experiment Description:   
(2000 characters)

Database Experiment Category:

Database Experiment SubCategory:

Green Channel (CH1) Description:

Red Channel (CH2) Description:

Reverse Replicate:

Normalization Type:   
If user defined, enter normalization value here:

- Print Name
- Slide Name
  - Unique, descriptive
- Experiment Date
- Experiment Name
  - Unique, descriptive
- Loading Prerequisites
- Channel Descriptions
- Reverse Replicate?
- Normalization Type
  - (describe later)

# Loading Data : Experiment Access

**EXPERIMENT ACCESS**

Database Experimenter: JCMATESE  
*Userid of person who performed experiment*

Enter group(s) and/or individual user(s) who should have permission to view this experiment.

Database Collaborative Group: (optional)  
Bialek Lab  
Boeke Lab  
Boothroyd Lab  
Botstein Lab  
Broach Lab

Database Individual User (select Database USERNAME): (optional)  
ECOX  
EKDE  
EMURPHY  
ENSMITH  
EPERLSTE

Check here if you want this experiment to be visible to all database users and the public.  
 NO  Yes

Load Experiment into Database Reset

- Experimenter (i.e. ‘Owner’)
  - Person who will have edit/delete/access privileges
- Collaborative Groups
  - By default, your lab group will be able to see all your experiments
  - If you wish for another entire group to view your experiments, you select the group name here
- Individual Users
  - Give an individual user the ability to view your experiment

# Loading Data : Experiment Access

**EXPERIMENT ACCESS**

Database Experimenter: JCMATESE  
*userid of person who performed experiment*

Enter group(s) and/or individual user(s) who should have permission to view this experiment.

Database Collaborative Group: (optional)  
Bialek Lab  
Boeke Lab  
Boothroyd Lab  
Botstein Lab  
Broach Lab

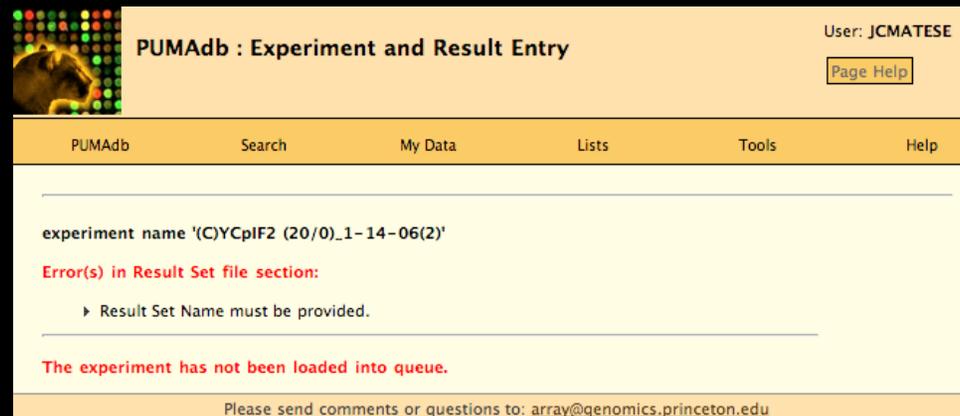
Database Individual User (select Database USERNAME): (optional)  
ECOX  
EKDE  
EMURPHY  
ENSMITH  
EPSON

Check here if you want this experiment to be visible to all database users and the public.  
 NO  Yes

Load Experiment into Database Reset

- World Access
  - Selecting ‘Yes’ here makes your data viewable by the WORLD
  - usually only done for open collaborations

# Loading Data : Errors



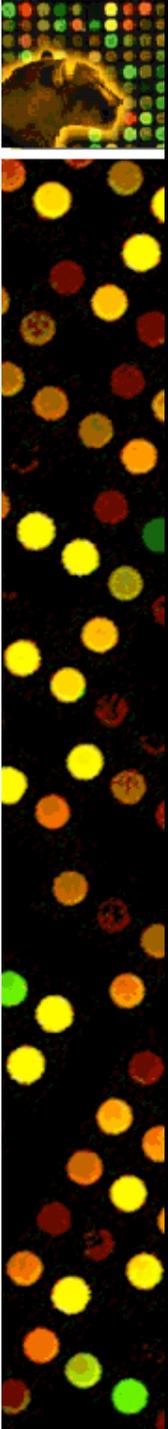
The screenshot shows the PUMAdb website interface. At the top, it says "PUMAdb : Experiment and Result Entry" and "User: JCMATESE". There is a "Page Help" button. Below this is a navigation bar with links for "PUMAdb", "Search", "My Data", "Lists", "Tools", and "Help". The main content area displays the experiment name "'(C)YcPIF2 (20/0)\_1-14-06(2)'" and a red error message: "Error(s) in Result Set file section: Result Set Name must be provided." Below the error message, it states "The experiment has not been loaded into queue." At the bottom, there is a footer with the text "Please send comments or questions to: [array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)".

- Loading software checks for common errors
- Experiments will not be loaded if there are errors. You must go back, correct your error(s) and resubmit your data



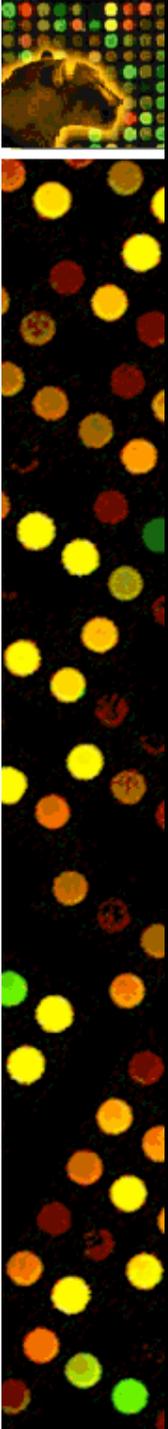
# Loading Data : Queue

- After passing the checks, your data goes into a loading *queue*
- The queue holds all experiments being loaded and processes them in an ordered fashion
- You can monitor the progress of your experiment entry
- You will also be sent an email with the hyperlink and Batch\_No to check the loading process



# Loading Data : Successful Experiment Entry

- Once your experiment has been loaded into the database, there are 2 methods to get the details of the experiment loading process
  - From the queue page
  - A file will be created on your loader account in the *logs* directory
    - `batch_no.log`



# Loading Data : Experiment Entry Log File

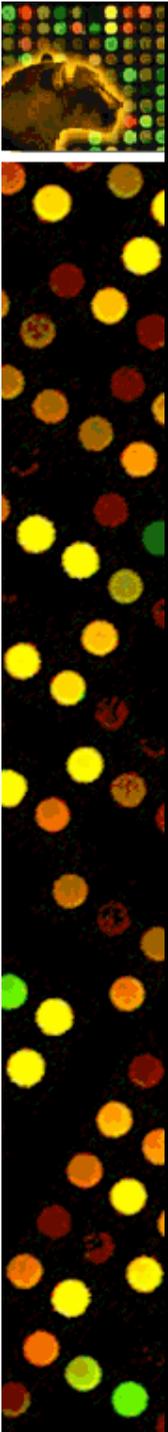
- The log file will give you the following information:
  - ExptID (experiment ID)
  - Information on experiment access
  - Information on normalization value
    - Number of spots that pass criteria
    - Spots used to calculate normalization
    - Percentage of spots that passed criteria
    - Normalization Value

# Loading Data : Batch Loading



The screenshot shows the PUMAdb website interface. At the top left is a logo featuring a mouse head and a grid of colorful dots. The main title is "PUMAdb : Experiment and Result Entry". In the top right corner, it says "User: UNRESTRICTED" and there is a "Page Help" button. Below the title is a navigation menu with links for "PUMAdb", "Search", "My Data", "Lists", "Tools", and "Help". The main content area contains three buttons: "Enter a New Experiment into the Database", "Enter a New Result Set for an Existing Experiment", and "Load Experiment(s) using a Batch File". The "Load Experiment(s) using a Batch File" button is highlighted with a yellow border.

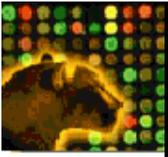
- Instead of loading experiments one by one, you can choose to load a batch of experiments
- All experiments need to be listed in a tab-delimited file (a batch file) in your *incoming* directory
- There are sample batch files located on the batch entry help page



# Loading Data : Assembling a Batch File

- (Result Set Name)
- Print Name
- Experiment Category
- Experiment SubCategory
- Slide Name
- Data File Location
- Grid File Location
- Green Scan File Location
- Red Scan File Location
- **Experiment Date**
- Experiment Name
- Green Channel (CH1) Description
- Red Channel (CH2) Description
- Normalization Type
- **Norm Value**
- Experimenter
- **Experiment Description**
- **Collaborative Group**
- **Individual User**

All underlined column headers are required data



# Loading Data via Batch File

	A	B	C	D
1	Print Name	Experiment Category	Experiment SubCategory	Slide Name
2	yeast6000-1	cell growth	glucose signaling	Gbc-051
3	yeast6000-1	cell growth	glucose signaling	Gbc-052

Print, Categorization, and Slide

E	F	G	H
Data File Location	Grid File Location	Green Scan File Location	Red Scan File Location
yw106_0-20_-03-15-02.gpr	proxy.gps	yw106_0-20_-03-15-02-W_532.tif	yw106_0-20_-03-15-02-W_635.tif
yw106_0-20_-10-8-02.gpr	proxy.gps	yw106_0-20_-10-8-02-W_532.tif	yw106_0-20_-10-8-02-W_635.tif

Filenames

I	J	K	L	M	N	O	Q	R
Experiment Name	Green Channel (CH1) Description	Red Channel (CH2) Description	Normalization Type	Norm Value	Experimenter	Experiment Date	Collaborative Group	Individual User
yw106_0-20_-03-15-02	Y2872 0min mRNA level	Y2872 20min mRNA level	user defined	0.841665	XIUyingz	11/28/01		
yw106_0-20_-10-08-02	Y2872 0min mRNA level	Y2872 20min mRNA level	user defined	0.860216	XIUyingz	8/16/02		

Experiment/Hybridization name, Experiment description, Normalization Type

Access

P
Experiment Description
Y2872 was grown in SC + 3% glycerol and shaken. When OD600 reached 0.64, cells were harvested as 0min, kept added 2% galactose to rest and kept growing for 20, 40, 60, 80 minutes. Lisa did all cell growth for this data batch.
Y2872 was grown in SC + 3% glycerol and shaken. When budding index reached 19%, cells were harvested as 0min, kept added 2% galactose to rest and kept growing for 20, 40, 60, 80 minutes



# Loading Data : Batch Loading

- After you select your organism, select your batch file
- First, check your batch file (catch common errors)...
- Next, queue/load your batch file...
- Load proceeds as for single experiment entry

Load Experiment(s) Using a Batch File

1. Choose Feature Extraction Software Package:

**New!** Data files can now be located within or in a subdirectory of:

- ▶ your incoming directory on loader
- ▶ your personal directory on arrayfiles

All files must be located in the same directory. Your batch file should **not** use any directory prefix. Please place the batch file in the same directory **OR** you can upload it from your own computer.

2. Choose the Directory Containing your Files:

On LOADER

or

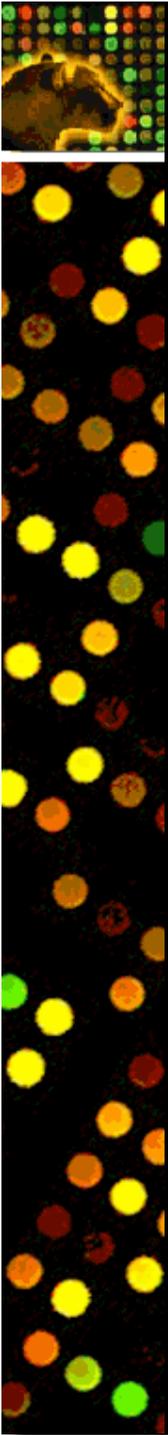
On ARRAYFILES

3. Choose a Batch File:

From the directory specified above

or

Upload from your computer   no file selected



# Loading Data : Example queue logfile

```
==== Loading Expt Batch NO : 3279 =====
```

```
Experiment Name: blah blah
```

```
Thu Dec 13 15:54:01 2001
```

```
Processing Data File : /loader/ftphome/youruserid/incoming/slidename.gpr
```

```
==== Inserting experiment info into experiment table... =====
```

```
exptID = 28765
```

```
The experiment data has been successfully inserted into experiment table!
```

```
==== Updating Experiment Access Control Table ... =====
```

```
Updating expt_access for experimenter YOURUSERID (#) ... OK
```

```
Updating expt_access for Brown/Botstein labs (#) ... OK
```

```
==== Calculate norm value... =====
```

```
Reading all data from datafile and doing all calculation now...
```

```
PassCriteria = 16005
```

```
Using 36490 spots for normalization 43.8% passed criteria of a good spot with 0.65
```

```
==== Updating exptNorm table... =====
```

```
NormType = Computed NormValue = 0.96
```

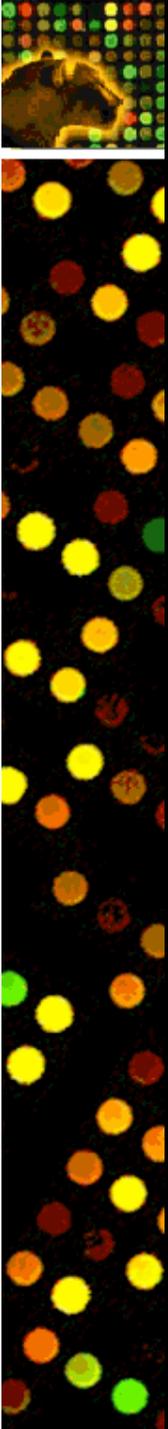
```
==== Updating Result table... =====
```

```
==== Total Record : 43200 =====
```

```
==== Updating Result table... =====
```

```
==== Expected = 43200, actual is 43200 =====
```

```
1000 . . .
```



# Replace a Proxy Image

## When:

- The image (.png) created by the default process is not acceptable
- After renormalization

## How:

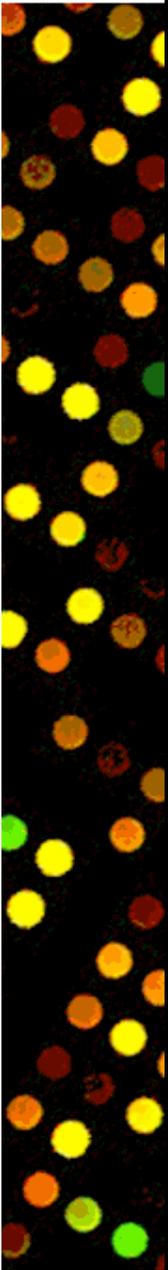
- Use your copy of tif files
- Make composite and save as .png
- Upload on loader into *incoming*
- Replace the copy
- Use the **Replace Proxy Image** link

Data Entry



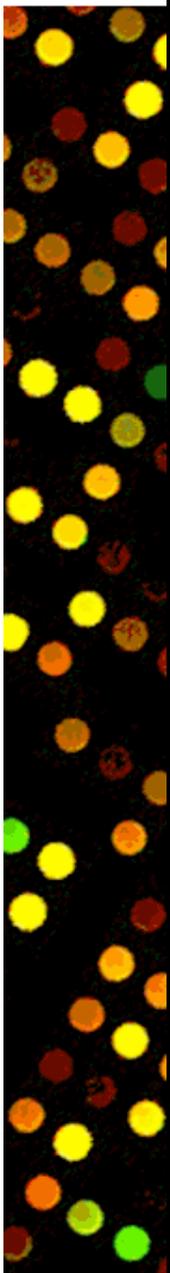
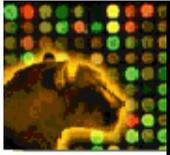
### EnterData

- › Experiments and Results
- › Replace Proxy Image
- › Patient and Clinical Data
- › Procedural Information
- › Protocol



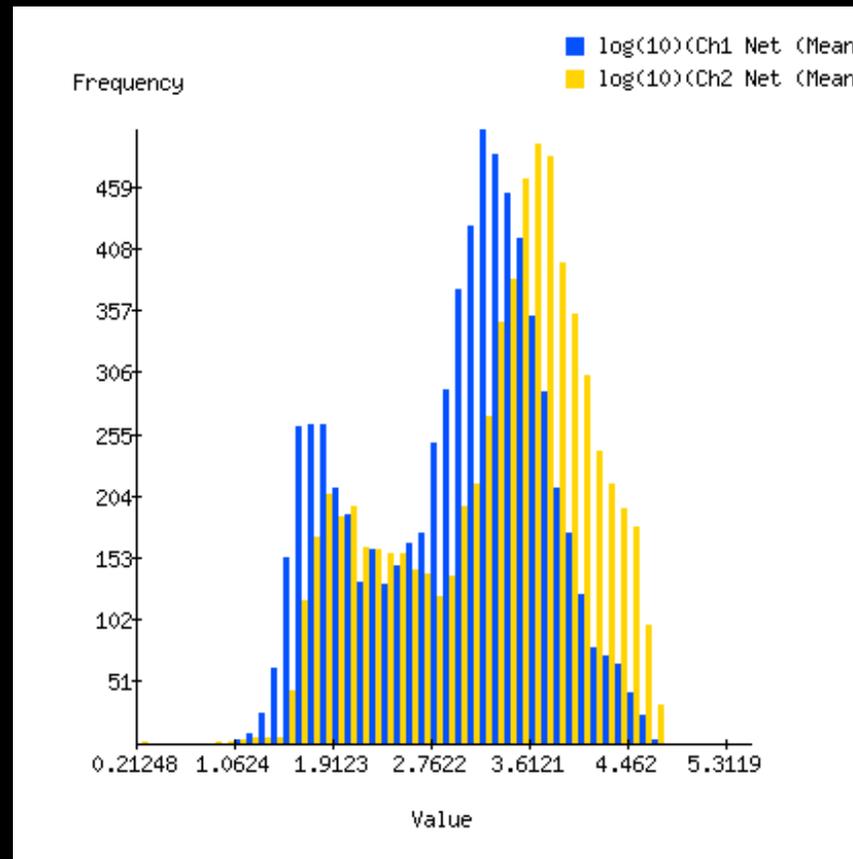
# Normalization: Why normalize data?

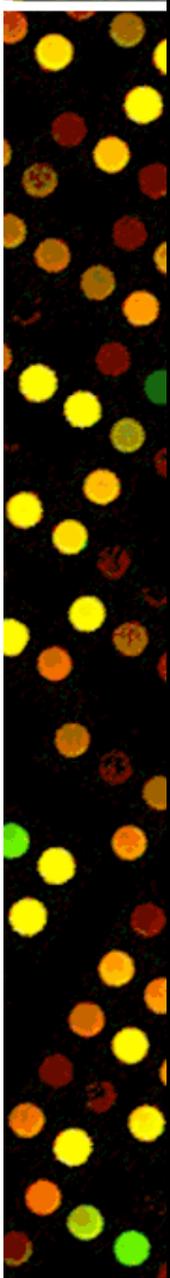
- Normalization reduces the effects of labeling bias
- Normalization allows you to recognize the biological information in your data
- Normalization allows you to compare data from one array to another



# Normalization: Channel biases

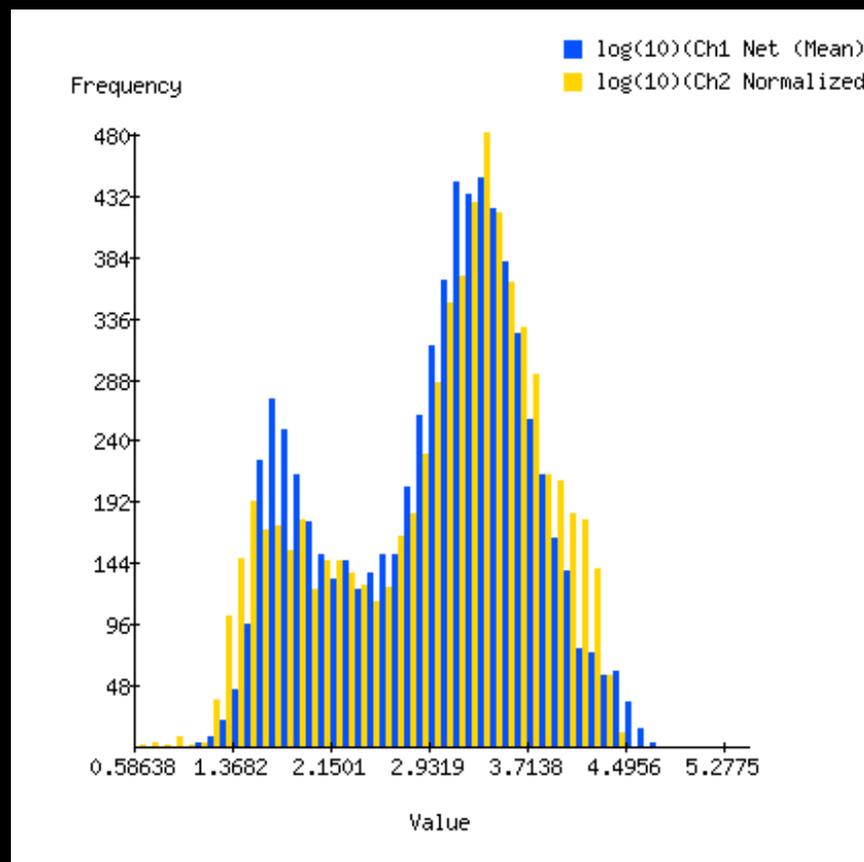
Before Normalization...

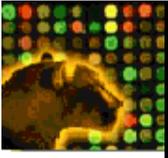




# Normalization: Channel biases

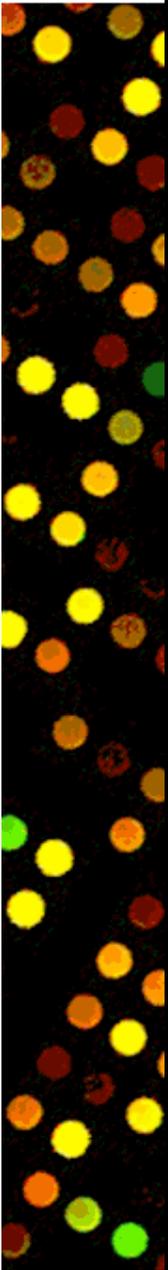
After Normalization...

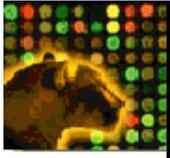




# Normalization Steps

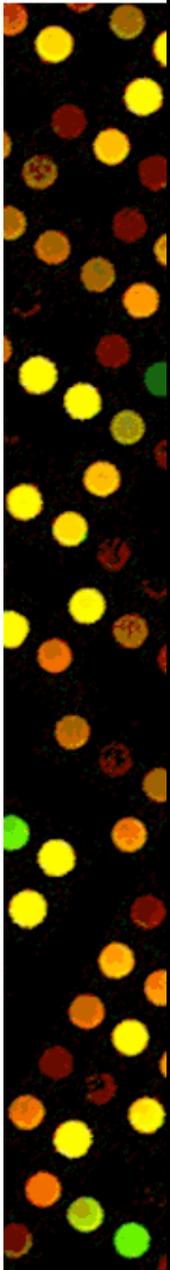
1. Assume that for the vast majority of spots on the array, the ratio should be 1 (i.e. no difference between samples/channels)
2. Choose those spots with “well-measured” data
3. Calculate a factor based on the initial assumption for these spots
4. Apply this factor to the second channel’s data for all spots

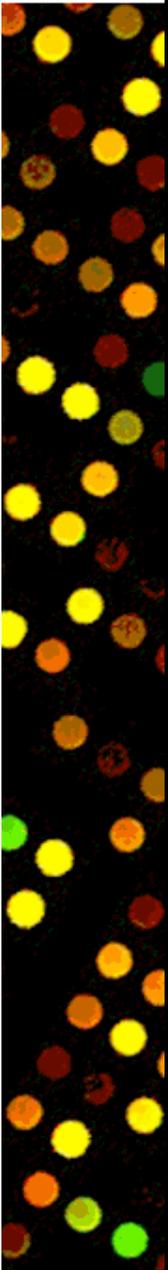
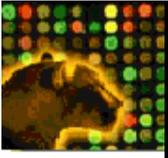




# Normalization: Choosing Spots

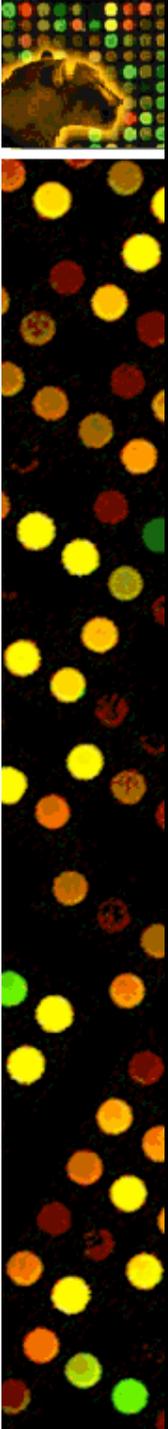
- The database offers two options for selecting “well-measured” spots for normalization:
  - “**Regression correlation**”: only non-flagged spots are used, with regression correlation greater than 0.6
  - “**Computed**”: based on the percentage of pixels in each unflagged spot whose intensity is at least one standard deviation greater than background (for Scanalyze spots, it is the fraction of pixels 1.5 fold greater than the background)





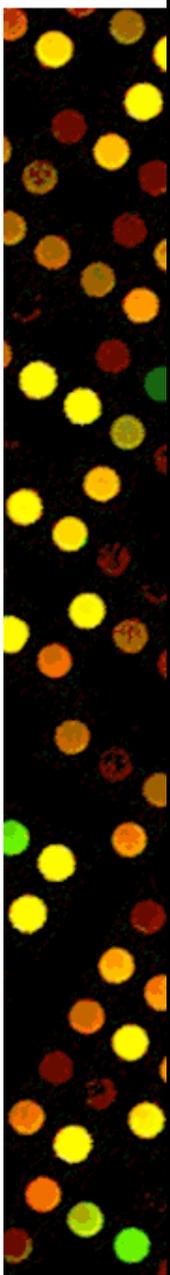
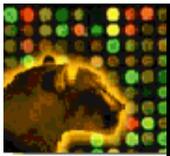
# Normalization: “computed” method

- “well-measured” spots are those with at least 65% of pixels significantly above background.
- If less than 10% of spots on the array meet the threshold, the 65% threshold is reduced stepwise until either 10% of spots pass or the threshold reaches 55% of pixels above background (whichever comes first)



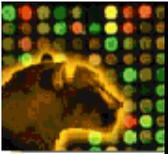
# Normalization: Calculating Factor

- Default normalization factor is the geometric mean of the red/green ratio of the selected “well-measured” spots
- Alternatively, a user can specify a normalization factor
- These methods can be applied for a genelist (in batch too)



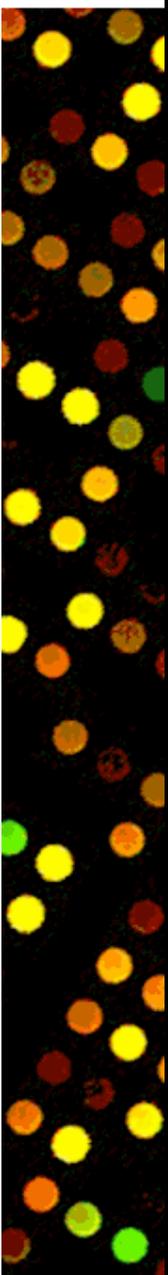
# Normalization: Applying the factor

- To apply the normalization factor, both the intensity and background of channel 2 (red) for all spots are divided by the normalization factor
- Other normalized values are calculated from these
- NOTE: Agilent data are not normalized in the database



# Welcome to PUMADB

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Data Retrieval and Analysis
- Organizing Data
- Submitting a Printlist

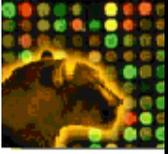


# Finding Your Data

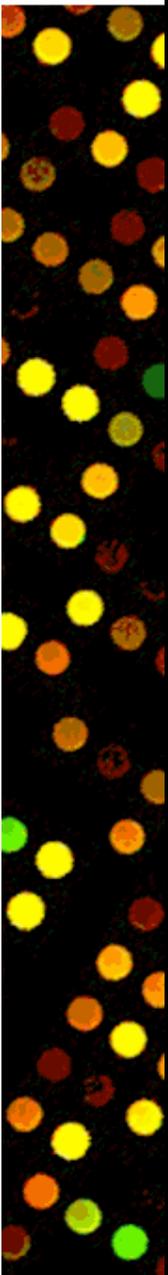


The screenshot displays the PUMAdd website interface. At the top, it says "PUMAdd : Index of Programs" and "User: UNRESTRICTED". The navigation menu includes "PUMAdd", "Search", "My Data", "Lists", "Tools", and "Help". The "Search" section is highlighted with a yellow box and contains links for "Advanced Search", "Basic Search", and "Gene Search". The "My Data" section is also highlighted with a yellow box and contains links for "Enter My Data", "Display My Data", "Select My Data", "Add Protocol", and "Procedural Info". The "Lists" section is highlighted with a yellow box and contains a list of links including "Users", "User Group", "Organism List", "Sequence Types", "SUID Retrieval", "Projects", "Plate Storage Locations", "Printers", "Print\_config", "Categories", "SubCategories", "Normalizations", "Plate List", "Print List", "Experiment List", "Procedures", "Protocol", "Parameters", "Experiment Types", "Experiment Set Types", "View Clinical Data", "Repository List", "Publication List", "Meta Data List", and "Ontology Browser". The "EnterData" section contains links for "Experiments and Results", "Replace Proxy Image", "Patient and Clinical Data", "Procedural Information", and "Protocol". At the bottom, it says "Please send comments or questions to: [array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)".

- There are several entry points for queries:
  - Advanced Search
  - Basic Search
  - Experiment List
  - Gene Search
  - Navigation Menu



# Finding Your Data : Basic Search



**Step 1: Pick results type**

<input type="radio"/> Publications	Browse published data by citation.
<input checked="" type="radio"/> My Experiment Sets	Browse organized experiment sets you created.
<input type="radio"/> Experiment Sets	Browse organized experiment sets by their name.
<input type="radio"/> Experiments	Browse viewable data by experimental category.

**Step 2: Browse selected results type**

First, choose an Organism	Second, choose Data Identifier
Caenorhabditis elegans Homo sapiens Plasmodium falciparum Rattus norvegicus Saccharomyces cerevisiae	Murphy et al. 2003 - All Murphy et al. 2003 - mutants Murphy et al. 2003 - RNAi1 Murphy et al. 2003 - RNAi2

There are three ways to find your data via Basic Search:

- Publications include all published data in the database
- Experiment sets allow you to search data on pre-defined experiment groups. (This will be described later)
- Retrieve data by experiment category

# Finding Your Data : Advanced Search Results

You may search for microarray experiment results using one of four methods:

- ▶ Method 1 – Allows you to select arrays by experimenter, category, subcategory and organism.
- ▶ Method 2 – Allows you to select experiments by printname.
- ▶ Method 3 – Allows you to use a premade list of arrays in your personal directory.
- ▶ Method 4 – Allows you to search for experiments using keyword(s).

*Be sure to click the radio button of the method you wish to use!*

Use Method 1 to select arrays by Experimenter, Category, Subcategory and/or Organism.

Select Organism:

Experimenter	Category	SubCategory
<input type="radio"/> And <input type="radio"/> Or	<input type="radio"/> And <input type="radio"/> Or	<input type="radio"/> And <input type="radio"/> Or
JABRISSO JACQUES JAGEE JALVAREZ JANA JASHAPIRO JASONLIH JCLIJU JCLORE JCMATESE	All Absolute transcript levels aCGH acute lymphoblastic leukemia Adenoma aging AKT Amino Acid metabolism Apoptosis Ascites	All 3'endseq 3-aminotriazole 4ats 4NQO AA_starvation aCGH cell line aCGH tumor tissue ACT1 Acute Myeloid Leukemia

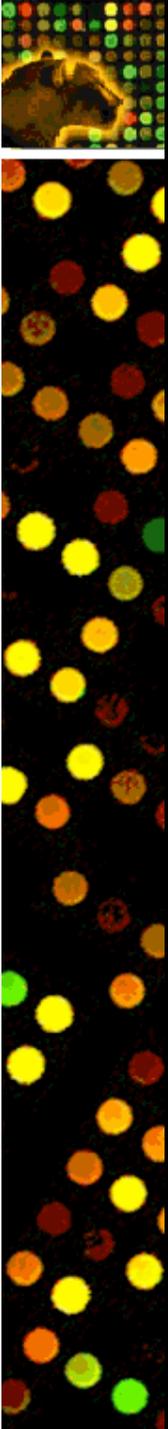
Use Method 2 to select all arrays from a given print:

Use Method 3 to use arrays from a list in your personal directory:

Use Method 4 to search experiment descriptions using keyword(s):

- ▶ Words should be separated by either '&' (and) or '|' (or), for example, 'carcinoma & renal'.

USING MY DEFAULTS ?



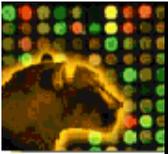
# Assay retrieval : Search software

## Use 'Basic Search' to browse/retrieve:

- a single Publication
- a single Experiment set
  - \* your personal sets
  - \* others', if viewable
- a single Experimental category

## Use 'Advanced Search' to perform:

- A boolean search
  - \* by Experimenter
  - \* by Category
  - \* by Subcategory
- A search by Print
- A search by arraylist
- A text search



# Welcome to PUMAdb

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Data Retrieval and Analysis
- Organizing Data
- Submitting a Printlist

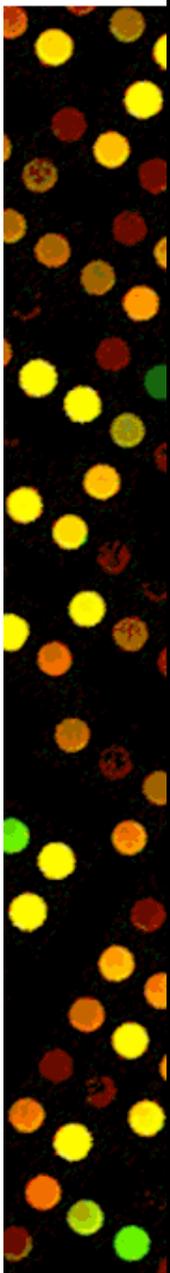
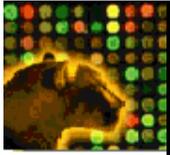
# Display Data from Searches

- ▶ Click on a Category/Subcategory to retrieve a list of all arrays using that category/subcategory
- ▶ Click on an experimenter to view their details
- ▶  = View and Sort Array Data
- ▶  = Download Raw Data
- ▶  = View Array Details
- ▶  = View Array Image and Grids
- ▶  = Clickable Image
- ▶  = Plot Array Data
- ▶  = Align Data to Chromosomes

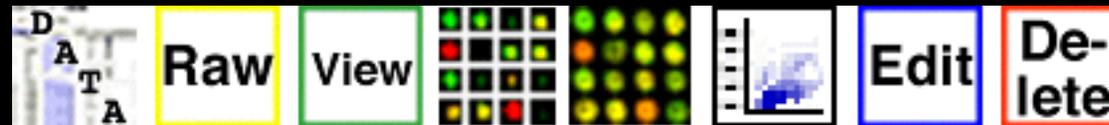
Your query returned 4 result sets.

Re-sort by:

ExptID	Experiment	Category	Subcategory	SlideName	Result Set	Options	Experimenter	ExptDate
101259	Y2864 2%glu (0-20)_6_15_05	<a href="#">Yeast expression</a>	<a href="#">glucose signaling</a>	251144713617_A01	RGT2-1	       	<a href="#">SZAMAN</a>	2005-06-15
101260	Y2864 2%glu (0-40)_6_16_05	<a href="#">Yeast expression</a>	<a href="#">glucose signaling</a>	251144713523_A01	RGT2-1	       	<a href="#">SZAMAN</a>	2005-06-16
101261	Y2864 2%glu (0-60)_6_15_05	<a href="#">Yeast expression</a>	<a href="#">glucose signaling</a>	251144713618_A01	RGT2-1	       	<a href="#">SZAMAN</a>	2005-06-15
101262	Y2864 2%glu (0-80)_6_15_05	<a href="#">Yeast expression</a>	<a href="#">glucose signaling</a>	251144713618_A02	RGT2-1	       	<a href="#">SZAMAN</a>	2005-06-15



# Display Data



- ◆  = **View and Sort Array Data**
- ◆  = **Download Raw Data**
- ◆  = **View Array Details**
- ◆  = **View Array Image and Grids**
- ◆  = **Clickable Image**
- ◆  = **Plot Array Data**
- ◆  = **Edit Array Details**
- ◆  = **Delete Array**

# Display Options: View Data

Sort by:

Display:

Display  rows, starting from row

Display NULLs Show results even if sorting field has no value.

Control spots: include control features.

Empty spots: include putatively empty features.

Select only features with no flag: include only features that have not been designated as unreliable either by the scanning software or by the array/hybridization owner.

Active Filter #	Measurement/Information	Operator	Value
<input checked="" type="checkbox"/> 1:	Regression Correlation	>	0.6
<input type="checkbox"/> 2:	Channel 1 Mean Intensity / Median Background Intensity	>	2.5
<input type="checkbox"/> 3:	Channel 2 Normalized (Mean Intensity / Median Background Intensity)	>	2.5
<input type="checkbox"/> 4:	Ch1 Net (Mean)	>=	350
<input type="checkbox"/> 5:	Ch2 Normalized Net (Mean)	>=	350
<input type="checkbox"/> 6:	Failed	=	0
<input type="checkbox"/> 7:	Is Contaminated	not equal	Y

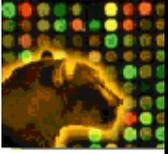
If you **do not** want the above criteria combined with a logical AND, enter a filter string (for example, "1 AND (2 OR 3)" or "1 AND ((2 OR 3) AND (4 OR 5)) OR 6").

Filter string:

- Select Data to use for sorting
- Select Columns to be displayed
  - Spot metrics
  - Biological Annotation
- Select how many rows to be displayed per page
- Include Controls/Nulls?
- Make downloadable file?
- Select filtering criteria

# Display Options: Raw Data

- This will save a file on your computer of all the columns of raw data
- The file is named *exptid.xls*
- The file is actually a tab-delimited file that can be opened in any program



View

# Display Options: View Details

Result set: RGT2-1	
Creator	SZAMAN
Result Set	794
Description	scanned in agilent scanner, feature extraction with all default settings
Data for : 251144713617_A01	
Experimenter	SZAMAN
Experiment ID	101259 <a href="#">Compare measurements with an experiment from the same Print</a>
Experiment Date	2005-06-15
Slide Name	251144713617_A01
Experiment Name	Y2864 2%glu (0-20)_6_15_05
Channel 1 Description	Y2864 in SC+3%glycerol to OD600=0.27 0min
Channel 2 Description	Y2864 in SC+3%gly+2%glu 20min
Is Reverse	N
Category	Yeast expression
Subcategory	glucose signaling
Normalization	normalization value is <b>undefined</b> <a href="#">Data Distribution</a>   <a href="#">Plot Data</a>   <a href="#">Signal Intensities</a>   <a href="#">Ratios on Array</a>
Print Information	
Print ID	2082 <a href="#">Generate GAL file for this print</a>
Print Name	Agilent-011447
Print Configuration	Agilent 1-sector
Description	
Description	<p>SZ249 strain has an integrated GAL10pRGT2-1 construct. Addition of galactose turns on expression of the dominant active RGT2-1 gene. RGT2 signaling is activated in the presence of high glucose. Therefore, comparison of galactose treated SZ249 cells to glucose treated wildtype cells should help identify the genes that are regulated by RGT2 in response to glucose.</p> <p>SZ249 and wildtype strains were grown in 3% glycerol media. Cells were collected at 0 min, treated with 2% galactose and then collected at 20, 40, 60 and 80 min post-treatment. In addition, wildtype strain was grown in 3% glycerol media and cells were collected at 0 min, treated with 2% glucose and collected at 20, 40, 60 and 80 min post-treatment.</p>
Experiment Data Sets	<a href="#">View Associated Experiment Sets</a>

- Gives you all the experimental annotation
- Allows you to compare measurements with another experiment from same print.
- Gives you the normalization method and value (if applicable)
- Gives you several options to access the quality of data



View

# Display Options: View Details

Normalization

Computed normalization value is 2.06

[Data Distribution](#)

[Plot Data](#)

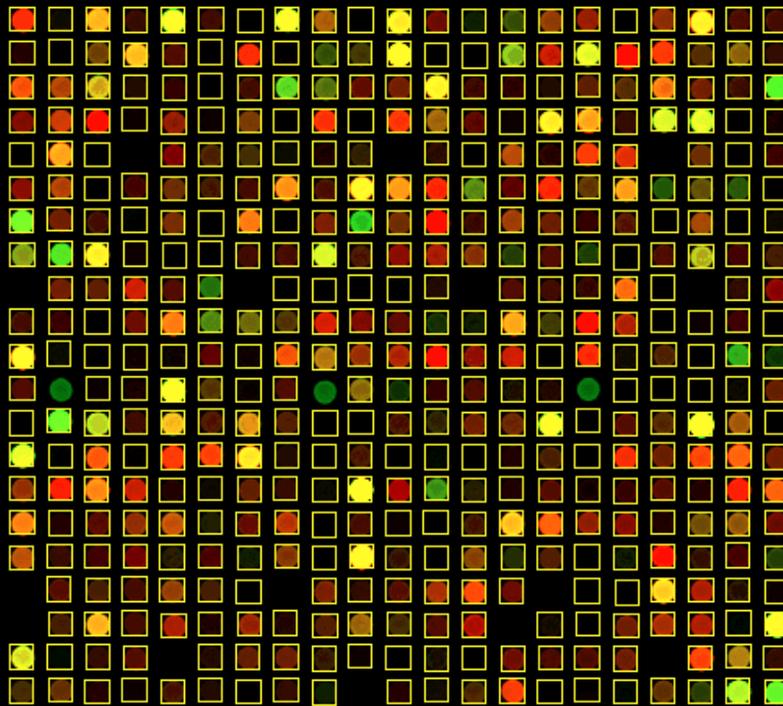
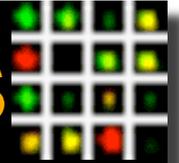
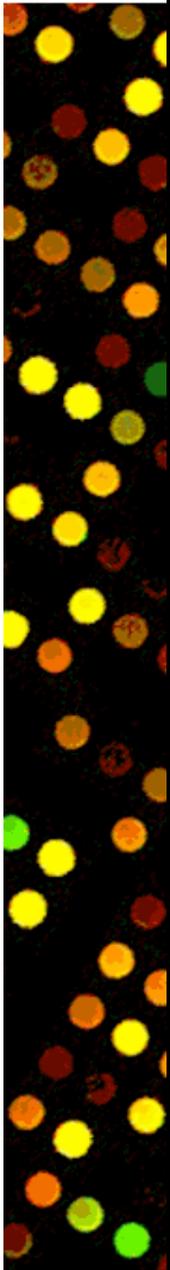
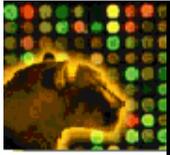
[Signal Intensities](#)

[Ratios on Array](#)

- Data Distribution
- Plot Data
- Signal Intensities
- Ratios on Array

*These graphs are covered in the data analysis tutorial.*

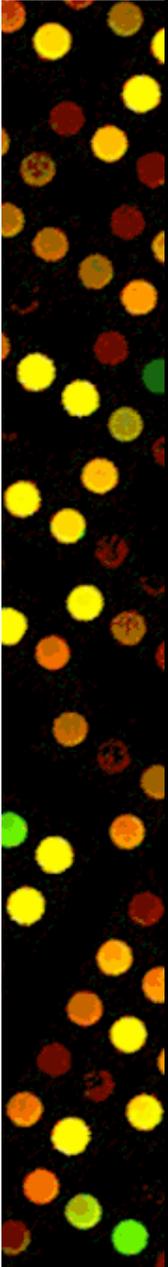
# Display Data: View images with grids



- Select data matching criteria
- Grid for array
- If you see a spot of interest, clicking on the spot yields...

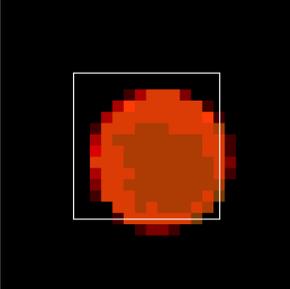


# Display Data: Spot Image



Individual Spot Data ( Agilent\_batch\_1)

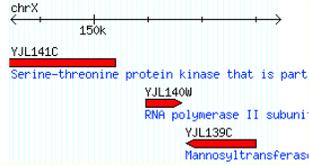
1 spot(s) found matching criteria.



Click spot to see in array context

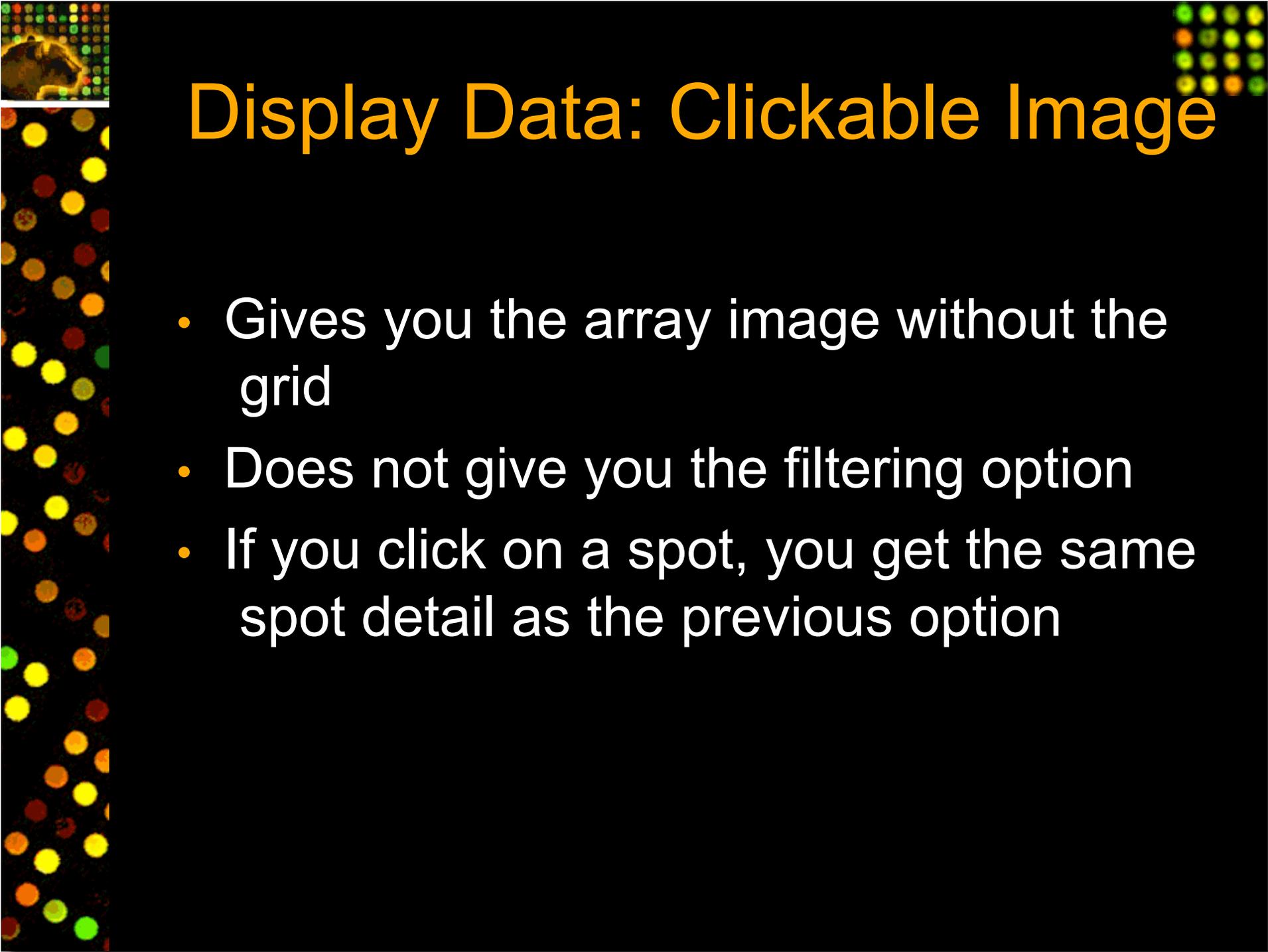
Biological Information	
Sequence Name	A_06_P4161
GENE NAME	BPB4
CHROMOSOME	10
STRAND	W
Beginning Coordinate	150958
Ending Coordinate	151623
BIOLOGICAL_PROCESS	transcription*
MOLECULAR_FUNCTION	transferase activity*
ORF Name	YJL140W
Sequence Description	YJL140W ebi.ac.uk:Database:sgd:A_06_P4161 www.chem.agilent.com:Database:agg

Relative Genome Map



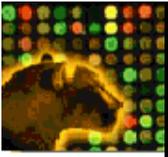
View expression history of this entity

Feature	Value	Flag Spot
X coordinate	746.908	
Y coordinate	1003.88	
Green Intensity (mean)	323.8644	
Green Intensity (median)	325	
Red Intensity (mean)	976.4407	
Red Intensity (median)	970	
Green Background (mean)	23.89744	
Green Background (median)	24	
Red Background (mean)	29.79121	
Red Background (median)	30	
Number of Green Inlier Pixels	59	
Number of Red Inlier Pixels	59	
Standard Deviation of Green Inlier Pixels	21.01515	
Standard Deviation of Red Inlier Pixels	57.41509	
Number of Green Background Pixels	273	
Number of Red Background Pixels	273	
Standard Deviation of Green Background Inlier Pixels	2.82916	
Standard Deviation of Red Background Inlier Pixels	4.219658	
Box Top	997	
Box Bottom	1010	
Box Left	740	
Box Right	753	
Feature Is Used for Global Background (0/1)		
log(base 10) (REDsignal/GREENsignal)	.1442506872	
Final Processed Green Intensity	4231.357	
Final Processed Red Intensity	5898.348	
GHighEndCorrDNSig		
RHighEndCorrDNSig		
Green Normalized Net Intensity	4231.36	
Red Normalized Net Intensity	5898.35	
Green Net Intensity	298.156	
Red Net Intensity	946.158	
Green Background Intensity Used	25.708	
Red Background Intensity Used	30.2825	
GProcessedSigError	38.82774	
RProcessedSigError	46.59793	
Error of Log Ratio	.05029484891	
log(base 10) (P-value of Log Ratio)	-2.38410865315161	
log(base 10) (P-value of Green Difference from Background)	-68.4480830206348	



# Display Data: Clickable Image

- Gives you the array image without the grid
- Does not give you the filtering option
- If you click on a spot, you get the same spot detail as the previous option



# Display Data: Plot Array Data

**Data Plotter Option Entry Form.**  
Choose the type of plot to make: Scatter Plot

Axis Plot?	Field	Scale	Log Base
X <input checked="" type="checkbox"/>	Final Processed Red Intensity	log	10
Y <input checked="" type="checkbox"/>	log(base 2) (REDSignal/GREENsignal)	linear	10

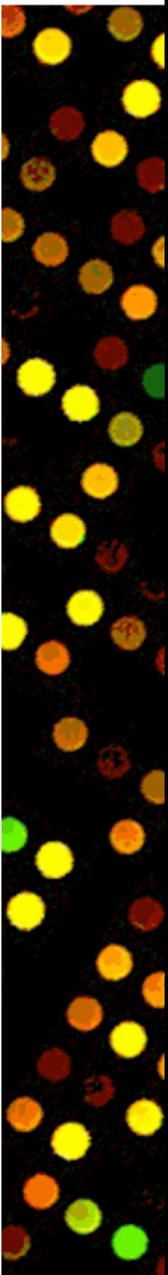
**Control spots:** include control features.  
 **Empty spots:** include putatively empty features.  
 **Select only features with no flag:** include only features that have not been designated as unreliable either by the scanning software or by the array/hybridization owner.

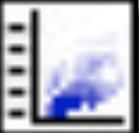
Active Filter #	Measurement/Information	Operator	Value
<input checked="" type="checkbox"/> 1:	Red Intensity Is Well Above Background (0 1)	=	1
<input checked="" type="checkbox"/> 2:	Green Intensity Is Well Above Background (0 1)	=	1
<input type="checkbox"/> 3:	Final Processed Red Intensity	>=	350
<input type="checkbox"/> 4:	Final Processed Green Intensity	>=	350
<input type="checkbox"/> 5:	Feature Is Red Population Outlier (0 1)	not equal	1
<input type="checkbox"/> 6:	Feature Is Green Population Outlier (0 1)	not equal	1
<input type="checkbox"/> 7:	Feature Is Red Non-uniformity Outlier (0 1)	not equal	1
<input type="checkbox"/> 8:	Feature Is Green Non-uniformity Outlier (0 1)	not equal	1

If you **do not** want the above criteria combined with a logical **AND**, enter a filter string (for example, "1 AND (2 OR 3)" or "1 AND ((2 OR 3) AND (4 OR 5)) OR 6").  
**Filter string:**

Display Reset

Evaluate data quality by plotting values for any array, using any measurement you wish to.



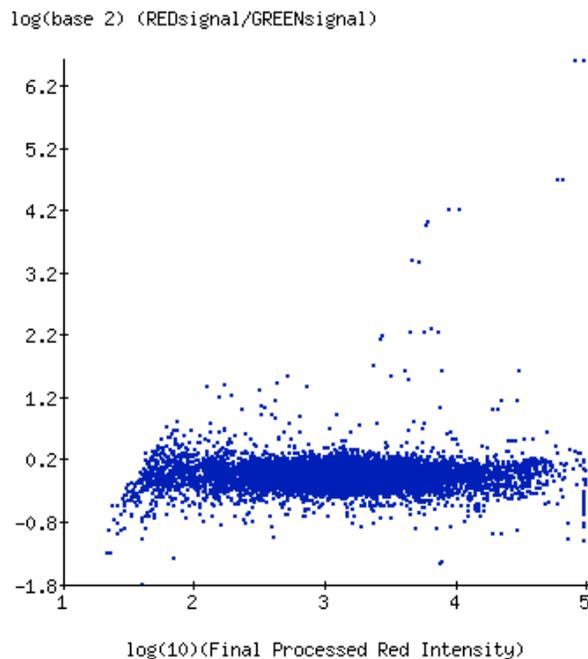


# Display Data: Plot Array Data

Plotting  $\log_{10}(\text{Final Processed Red Intensity})$  vs.  $\log_{\text{base } 2}(\text{REDSignal}/\text{GREENsignal})$

Features with values that could not be log-transformed have been omitted from the plot(s).

251144713619\_A02 (Y2864 2%gal (0-80)\_6\_15\_05 - RGT2-1)  
9496 features with good data



Evaluate data quality by plotting values for any array, using any measurement you wish to.

# Display Data: Edit Experiment Details

**Data for : test rlu load**

Slide Name: test rlu load

Experiment Name: test rlu load

Experiment Date: 2005-04-13 (YYYY-MM-DD)

Channel 1 Description: test rlu load

Channel 2 Description: test rlu load

Reverse Replicate: N

Clinical Data: No Patient Information No Clinical Information

Experiment Type: **Choose One** [Submit Changes](#)

Procedural Information: *Procedural information parameters have been entered before associating or editing procedural information.*

Category: Chromatin IP

Subcategory: Expression (Type I)

Current Normalization: Gene Expression (single channel) 1.58 [View Data Distribution](#)

Renormalize: [Select normalization options](#)

Add Access

Groups: AFGC, Amon Lab, Aphid Consortium, Arvin Lab, Barsh lab

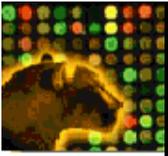
Users: AACUPTA, AARNOLD, AAZVOLIN, ABRUKMAN, ACCERBI

- Edit all names and descriptions
- Experiment Type
- Associate procedural information
- View Data Distribution
- Re-normalize data

# Display Data : Editing Access

Remove Access	Groups	Broach Lab	Users	FKANG
	Groups	AFGC Amon Lab Aphid Consortium Arvin Lab Barsh lab	Users	AAGUPTA AARNOLD AAZVOLIN ABRUKMAN ACCERBI

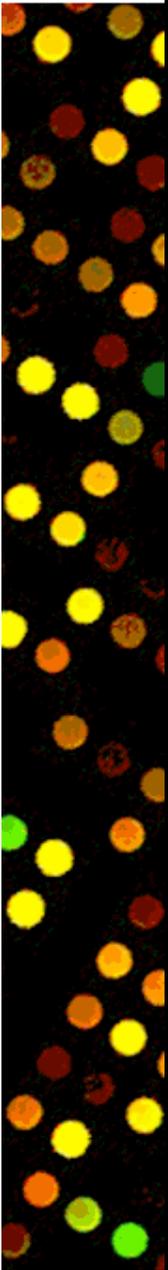
- Under Edit Experiment Details, you can add or remove experiment access
- You can give access to an entire group or an individual user
- To give access to collaborators:
  - Register your collaborator
  - In Experiment Details, Click on collaborator's name to grant access to view experiment

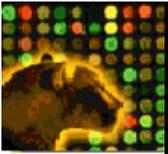


De-  
lete

# Display Data: Delete an Experiment

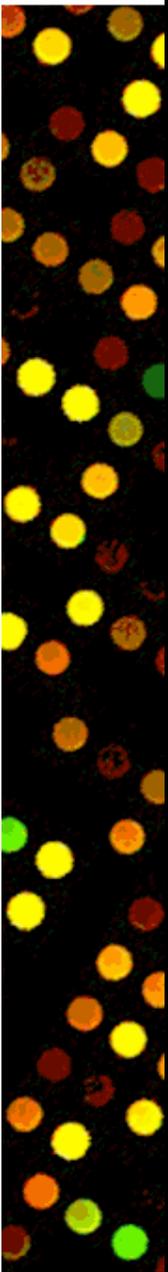
- Only the owner of an experiment can delete it
- Once an experiment is deleted from the database, it can not be recovered easily
- Once an experimenter leaves the lab, the lab head should consider what to do with his/her experiments, i.e. should the user still have the ability to delete all their experiments?





# Welcome to PUMAdb

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Organizing Data
- Submitting a Printlist



# Organizing Data: Data Retrieval and Analysis

Your query returned **76** result sets.  
Select Experiment Names from the following List:

```
251144713617_A01 :: Y2864 2%glu (0-20)_6_15_05 (AGILENT - RGT2-1)
251144713523_A01 :: Y2864 2%glu (0-40)_6_16_05 (AGILENT - RGT2-1)
251144713618_A01 :: Y2864 2%glu (0-60)_6_15_05 (AGILENT - RGT2-1)
251144713618_A02 :: Y2864 2%glu (0-80)_6_15_05 (AGILENT - RGT2-1)
251144713620_A01 :: Y2864 2%gal (0-20)_6_15_05 (AGILENT - RGT2-1)
251144713620_A02 :: Y2864 2%gal (0-40)_6_15_05 (AGILENT - RGT2-1)
251144713619_A01 :: Y2864 2%gal (0-60)_6_15_05 (AGILENT - RGT2-1)
251144713619_A02 :: Y2864 2%gal (0-80)_6_15_05 (AGILENT - RGT2-1)
251338410236_A01 :: Y2866 2%gal (0-20)_12-7-05 (AGILENT - GAL10pRAS2V19_12_7_05)
251338410237_A01 :: Y2866 2%gal (0-40)_12-7-05 (AGILENT - GAL10pRAS2V19_12_7_07)
251338410237_A02 :: Y2866 2%gal (0-60)_12-7-05 (AGILENT - GAL10pRAS2V19_12_7_08)
251338410236_A02 :: Y2866 2%gal (0-80)_12-7-05 (AGILENT - GAL10pRAS2V19_12_7_06)
251144712827_01 :: WT -60/0 (AGILENT - RAS dominant negative)
251144712827 :: WT -60/20 (AGILENT - WT -60/20)
251144713120_01 :: WT -60/40 (AGILENT - RAS dominant negative)
251144713121_02 :: WT -60/60 (AGILENT - RAS dominant negative)
251144713122_02 :: Ras N24 -60/ WT -60 (AGILENT - RAS dominant negative)
251144713122_A01 :: Ras N24 -60/+60 reloaded (AGILENT - RAS dominant negative)
251144712828_01 :: Ras N24 -60/0 (AGILENT - RAS dominant negative)
251144713121_01 :: Ras N24 -60/20 (AGILENT - RAS dominant negative)
251144713121_02 :: Ras N24 -60/40 (AGILENT - RAS dominant negative)
251144714253_A01 :: S2268 2%gal (0-20)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714253_A02 :: S2268 2%gal (0-40)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714287_A01 :: S2268 2%gal (0-60)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714287_A02 :: S2268 2%gal (0-80)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714288_A01 :: S2268 2%glu (0-20)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714288_A02 :: S2268 2%glu (0-40)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714282_A01 :: S2268 2%glu (0-60)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714282_A02 :: S2268 2%glu (0-80)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
251144714251_A01 :: S2271 2%glu (0-20)_11-16-05 (AGILENT - GAL1pSCH9_11_16_05)
```

Display Data  
Data Retrieval and Analysis  
Create Result Set List  
Make an Experiment Set  
Reset

- Once you have selected a group of experiments, you need to select the experiments you wish to work with
- You have several different options:
  - *Display Data*
  - *Data Retrieval and Analysis (clustering)*
  - *Create Result Set List*
  - *Create Experiment Set*

# Organizing Data: Result Set List

## Result Set List Creation

[Contents and Template](#) -> Customization

Enter a name for your ResultSetList:  (mandatory)

Organize Your Result Sets:

Starting List of Result Sets	Result Sets Included within ResultSetList
<p>Y2864 2%gal (0-20).6_15_05 (AGILENT - RGT2-1)            Y2864 2%gal (0-40).6_15_05 (AGILENT - RGT2-1)            Y2864 2%gal (0-60).6_15_05 (AGILENT - RGT2-1)            Y2864 2%gal (0-80).6_15_05 (AGILENT - RGT2-1)</p> <p>&gt; Add &gt;            &lt; Remove &lt;            &lt;=&gt; Add All &gt;&gt;            &lt;&lt; Remove All &lt;&lt;</p> <p>Sort Asc   Sort Desc</p>	<p>Y2864 2%glu (0-20).6_15_05 (AGILENT - RGT2-1)  <b>Y2864 2%glu (0-40).6_16_05 (AGILENT - RGT2-1)</b>            Y2864 2%glu (0-60).6_15_05 (AGILENT - RGT2-1)            Y2864 2%glu (0-80).6_15_05 (AGILENT - RGT2-1)</p> <p>Sort Asc   Sort Desc</p>

**Organize the arraylist, if desired.**

Move Experiment(s) Up  
 Move Experiment(s) Down

Select **Template for Filters**: these may be customized for each result set on the next page or by editing the file.

› Select default filters for Agilent feature extraction software data

Active Filter #	Measurement/Information	Operator	Value
<input type="checkbox"/> 1:	Red Intensity Is Well Above Background (0 1)	=	1
<input type="checkbox"/> 2:	Green Intensity Is Well Above Background (0 1)	=	1
<input type="checkbox"/> 3:	Final Processed Red Intensity	>=	350
<input type="checkbox"/> 4:	Final Processed Green Intensity	>=	350
<input type="checkbox"/> 5:	Feature Is Red Population Outlier (0 1)	not equal	1
<input type="checkbox"/> 6:	Feature Is Green Population Outlier (0 1)	not equal	1
<input type="checkbox"/> 7:	Feature Is Red Non-uniformity Outlier (0 1)	not equal	1
<input type="checkbox"/> 8:	Feature Is Green Non-uniformity Outlier (0 1)	not equal	1

If you **do not** want the above criteria combined with a logical **AND**, enter a filter string (for example, "1 AND (2 OR 3)" or "1 AND ((2 OR 3) AND (4 OR 5)) OR 6").

Filter string:

[Customize Filters](#)   [Create ResultSetList](#)

- › 'Customize Filters' allows you to modify the filtering parameters above for every result set in the list (if the list contains no more than 20 result sets).
- › 'Create ResultSetList' skips the customization step, and creates the ResultSetList directly, with any active filters.

# Organizing Data: Experiment Sets

Experiment Set Name:

Experimental Set Design:

Experimental Set Longevity:  
 Permanent set  
 Temporary workset (removed in 30 days)

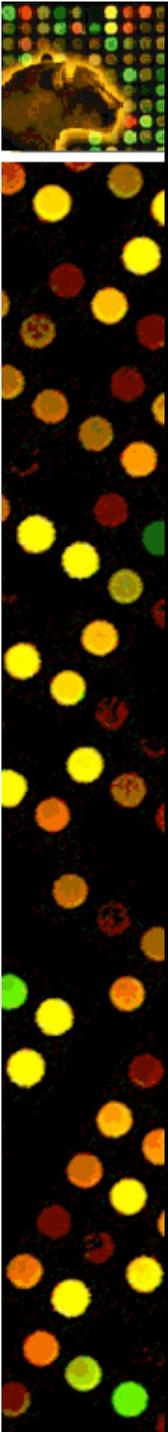
Experiment Name	Reverse	Cluster Weight	Result Set Name	Experimental Factors and Values
				Time Description:
Y2864 2%glu (0-20)_6_15_05 (AGILENT) N	N	<input type="text" value="1"/>	RCT2-1	<input type="text" value="20"/> minutes
Y2864 2%glu (0-40)_6_16_05 (AGILENT) N	N	<input type="text" value="1"/>	RCT2-1	<input type="text" value="40"/> minutes
Y2864 2%glu (0-60)_6_15_05 (AGILENT) N	N	<input type="text" value="1"/>	RCT2-1	<input type="text" value="60"/> minutes
Y2864 2%glu (0-80)_6_15_05 (AGILENT) N	N	<input type="text" value="1"/>	RCT2-1	<input type="text" value="80"/> minutes

Experiment Set Description (Detailed as possible):

Do you want to publish this experiment set?  NO  YES

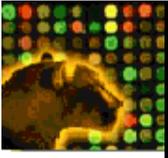
Create Experiment Set

- Order your experiments
- Select experimental factors (optional)
- Next provide more details
  - Name, Experiment set design, Longevity
  - Weights for clustering
  - Set description
    - For publications, this would be the abstract or figure legend
  - Publication Radio Buttons
    - All experiments must be world viewable in order to publish the set



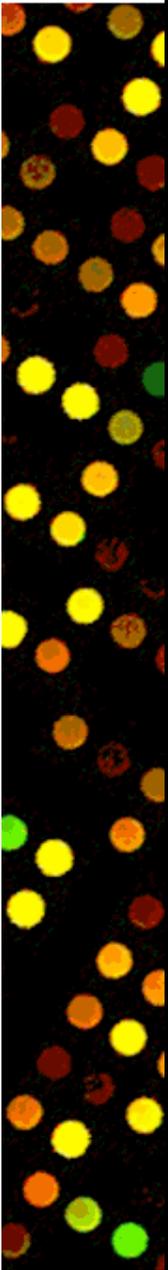
# Organizing Data: Result Set List vs Experiment Set

- Result Set (Arraylist)
  - Text file that exists in your loader account arraylists directory
  - Only visible to you
  - Contains no annotation
  - Customized filtering
  - Accessed through Advanced Search
- Experiment Set
  - Exists in the database therefore dynamic (edit, delete, or annotate through a web interface)
  - Visible to users & collaborators
  - Can be well annotated
  - Required for publication within the database
  - Accessed through Basic Search



# Organizing Data: Genelists

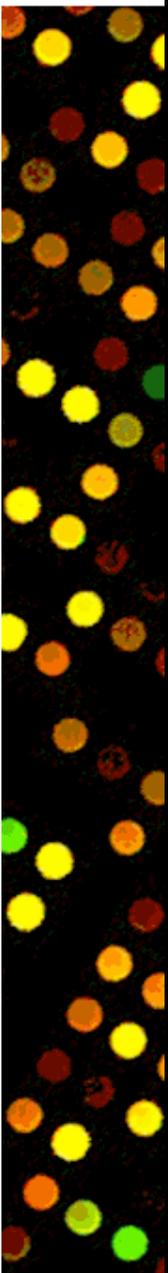
- What is a genelists?
  - A file containing a list of genes that exists in your loader account in the directory *genelists*
- What is the purpose of a genelists?
  - Cluster and analyze only a set of genes
  - When retrieving your data, you may choose to retain the annotation from your genelists instead of using the database annotation
- There are several shared standard files of genelists that are available for many organisms.
- You may create your own precompiled list of genes.
- Normalization values can be calculated based on a genelists.





# Organizing Data: Creating your own genelists

- Create a tab-delimited text file
- The first line of the file must have the appropriate label for the data contained within it
  - NAME (YPR119W, IMAGE:1542757, or HPY1808)
  - SUID
  - LUID
  - SPOT
- Your file may contain one additional column with any type of annotation data you desire for each gene
- This information can be extracted during data analysis and carried all the way over through clustering





# Questions?

Send e-mail:

[array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)

Office: CIL 135

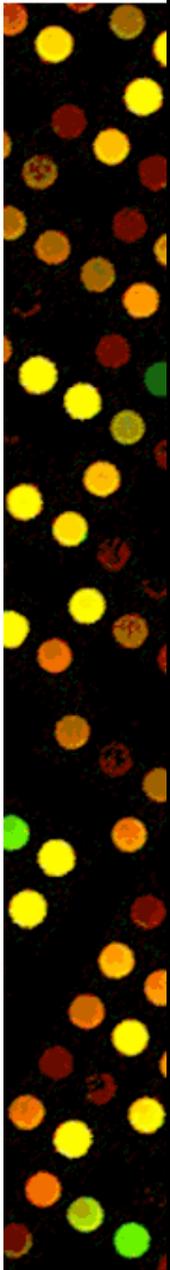
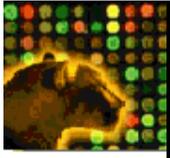
Phone: 258 - 8309

Online help: <http://puma.princeton.edu/help/>



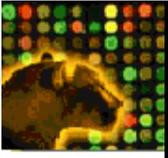
# Welcome to PUMADB

- User Registration
- Staging Data
- Loading Prerequisites
- Loading Data
- Finding Your Data
- Displaying Your Data
- Data Retrieval and Analysis
- Organizing Data
- Submitting a Plate Samples



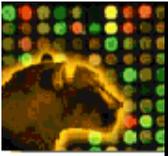
# Submitting Plate Samples and ArrayDesigns

- The creation of a print within the database is a complex process.
- If you receive your arrays from the core facility, this is done for you
- Plate samples are conveyed as a tab-delimited list (well address + contents)
- There is a program to assist you in platesample submission:
  - Located under “*Tools*” on the “**Index of Programs**” page
  - Printlist must be in your incoming directory on loader



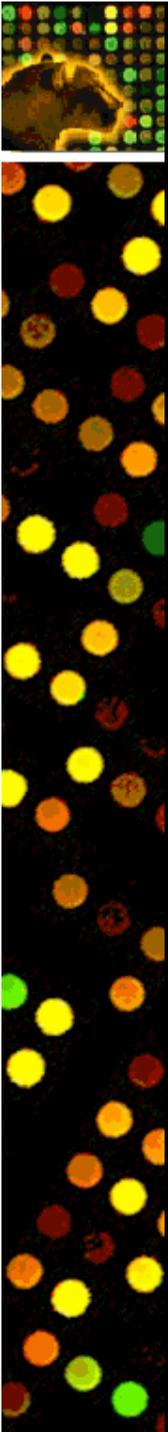
# Submitting Plate Samples : Is a new list required?

- **Yes**, if the plates used have not been previously entered into the database
- **Yes**, if the plate was entered in the past, but their contents have changed over time (well contamination, well emptied)
- **No**, if your lab makes 3 different prints using the exact same plates in the same or different order
  - Just need to tell a curator the a list of database plateIDs and plateNames from the first print in their new order.



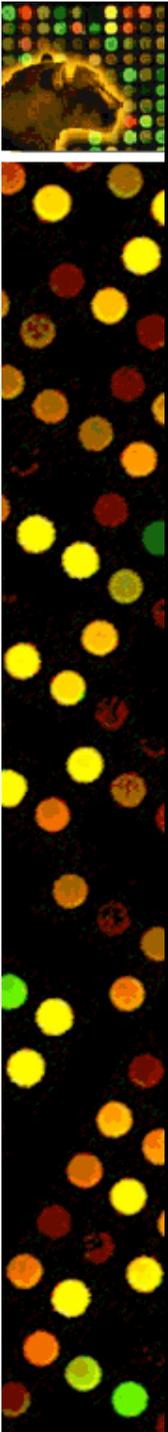
# Submitting Plate Samples: Column Headers

- **PLAT**: The plate number; eg 1, 2, 3, etc. \***INTEGER**\*
- **PROW**: The plate row; eg A, B, C, etc. \***CHARACTER**\*
- **PCOL**: The plate column; eg 1, 2, 3, etc. \***INTEGER**\*
- **NAME**: The sequence name
  - usually a systematic name or clone identifier (I.e. YBL016 or IMAGE:753234)
  - This is the only name used for samples of TYPE other than CDNA.
- **TYPE**: The sequence type
  - Usually ORF, CDNA, CONTROL, or EMPTY.
  - List of types can be seen from the SMD homepage under *List Data: Sequence Type*
- **FAIL**: Whether the PCR failed
  - 0 : one distinct band - success
  - 1 : no signal - fail
  - 2 : multiple distinct bands
  - 3 : signal, but not a distinct band (smear)
  - 4 : multiple smears
  - 5 : unknown
  - 101 : worst cases of peeled away or haloed spots(assigned on a 96 well plate basis)
  - 102 : less bad cases of peeled away or haloed spots(assigned on a 96 well plate basis)
  - Null is assumed to be 0 (success)



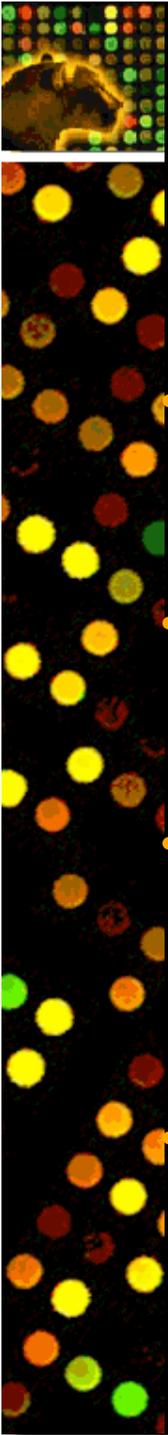
# Submitting Plate Samples: Additional Columns for cDNA data

- **CLONEID**:
- **ACC**: Required if CLONEID is absent/null.
  - This is the GenBank accession, usually acquired from dbEST.
- **IS\_CONT**: Whether the sample is known to be contaminated. A blank entry will default to unknown (U)
- **IS\_VER**: Whether the DNA in a well has been verified. A blank entry will default to unverified (U).
- **SOURCE**: A string describing the source of the clone or DNA. This has typically been used to indicate the original plate source, and the 96 and 384 well plate locations that a clone has been in
  - GF200:96(1A1):384(1A1).
  - GF200 refers to a set of resgen plates



# Submitting Plate Samples: Optional Columns

- **DESC**: A description of the molecular entity. This description is associated with the SUID itself (not a clone or platesample description)
- **LUID**: Laboratory Unique ID: For those samples that have identical NAME and TYPE, but require distinction within the laboratory for experimental reasons (different sources, new PCR, new plate). If you wish to enter LUIDs for your lab's platesamples, please contact the curators: [array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)
- **GENE\_NAME**: Sometimes clones will stop being included in UniGene for spurious reasons, but users have a 'Preferred Name' for those clones.
- **ORIGIN**: For CDNA clones, this can indicate whether this is a public or private clone.
- **SAMPLE\_DESC**: A description, if any, about that particular sample. This description is specific to the plate sample.
- **ORGANISM**: If submitting a print containing samples from multiple organisms (i.e. human, yeast). For those few rows where the sample is derived from an organism \*other\* than the default (user-defined), the organism code must be specified.



# Submitting Plate Samples: Creating New SUIDs

New samples in your plates (i.e. those not currently in the database) will need to have a unique sequence identifier assigned to them (SUID)

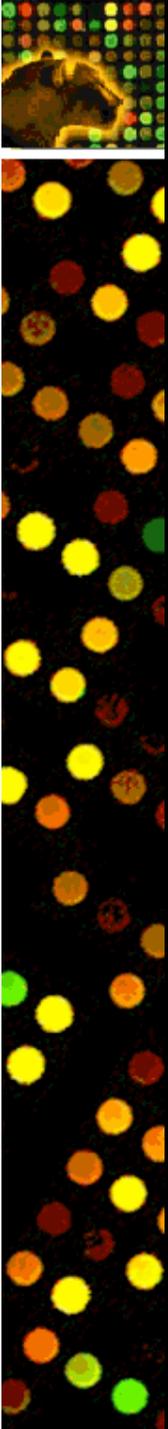
A SUID is meant to represent a unique molecular entity within the database. It is relatively meaningless outside the context of the database.

The combination **NAME::TYPE::ORGANISM** uniquely identify an SUID

- YBL001C::ORF::SC → SUID:3429
- IMAGE:486544::CDNA::HS → SUID:28546

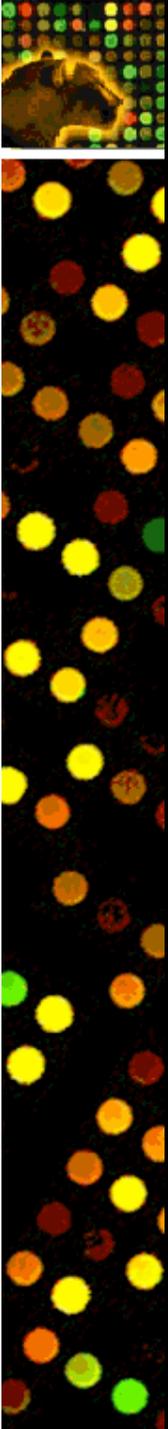
SUIDs allow comparison of the same samples across different prints.

- It is extremely important that erroneous SUIDs are not created.
  - This will prevent comparisons between prints/experiments



# Submitting Plate Samples: Avoiding Common Name Errors

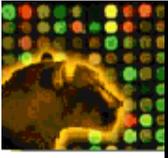
- Erroneous SUIDs are usually created by a bad NAME
  - misspelled, non-standard, or non-systematic
    - ACT1:ORF:SC or Actin:ORF:SC → YFL039C:ORF:SC
    - 3X SSC:CONTROL:SC → 3xssc:CONTROL:SC
  - Every new sample must be verified by the user before it is assigned a new SUID and before the printlist can be entered.
  - Please be a conscientious user and verify that any new SUIDs you approve are valid.
- Empty wells must be specified as such
  - All empty wells must be designated NAME=>EMPTY and TYPE=>EMPTY.
  - Do not use "blank" or "control" to describe empty wells.



# Submitting a Printlist: Avoiding Common Errors

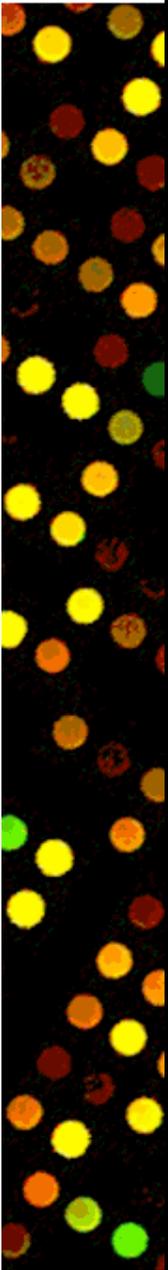
- Headers misspelled or absent
- Required data missing
  - except FAIL, CLONEID, but column header must still be present
- Correct Plate ordering
- No wells may be skipped (with the exception of the last plate in the print run).
- Useful check: number of plate samples = number of printed spots

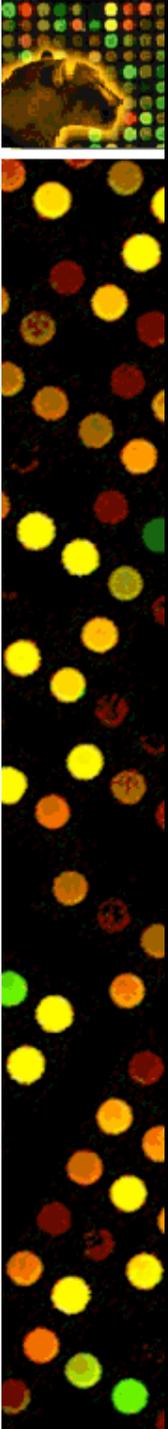
$\#samples = (\#printlist\ rows - 1) \leq \#tips * \#rows\ per\ sector * \#columns\ per\ sector = \#spots$



# Submitting Plate Samples: Validation Program

- The printlist must be placed in your *incoming* directory on your loader account
- This program will assist you in printlist submission
  - It follows the rules stipulated above.
- The program will send all feedback to your *logs* directory
  - Filename.new
  - Filename.errors





# Submitting a Printlist: Notify Curators

- Additional information needed:
  - Number of sector rows/columns
  - Distance of rows/columns in sector
  - Printing algorithm: <http://puma.princeton.edu/help/createPrint.shtml>
  - Number of slides printed
  - Plate location
  - Printer used for printing
- When your printlist is correct - send email with info above to [array@genomics.princeton.edu](mailto:array@genomics.princeton.edu)