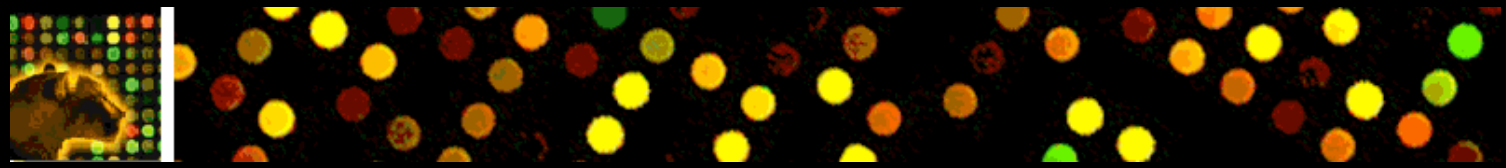
Database Repository and Tools



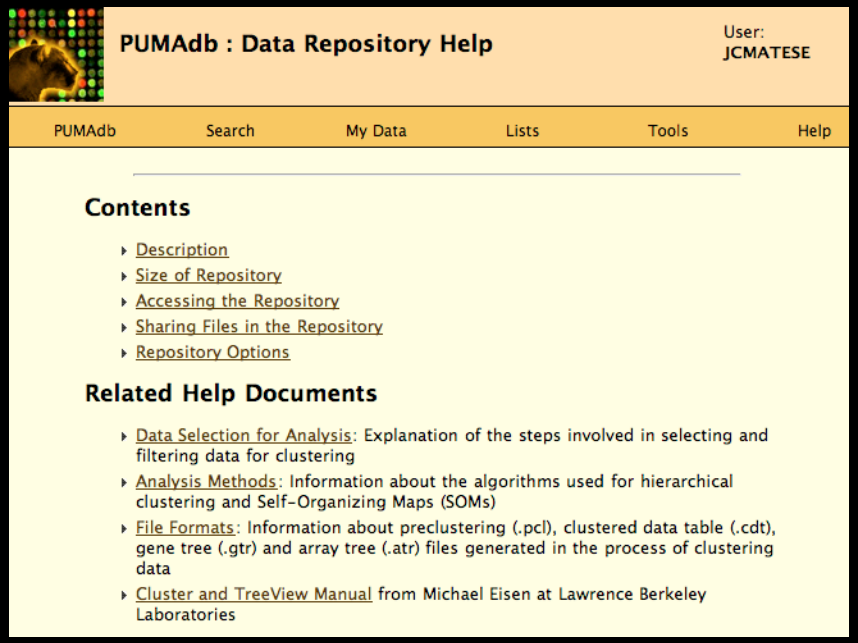
John Matese

May 9, 2008

What is the “Repository”?

- Save and exchange retrieved and analyzed **datafiles**
- Perform datafile manipulations (averaging and annotations)
- Run specialized algorithms (imputation, SVD)

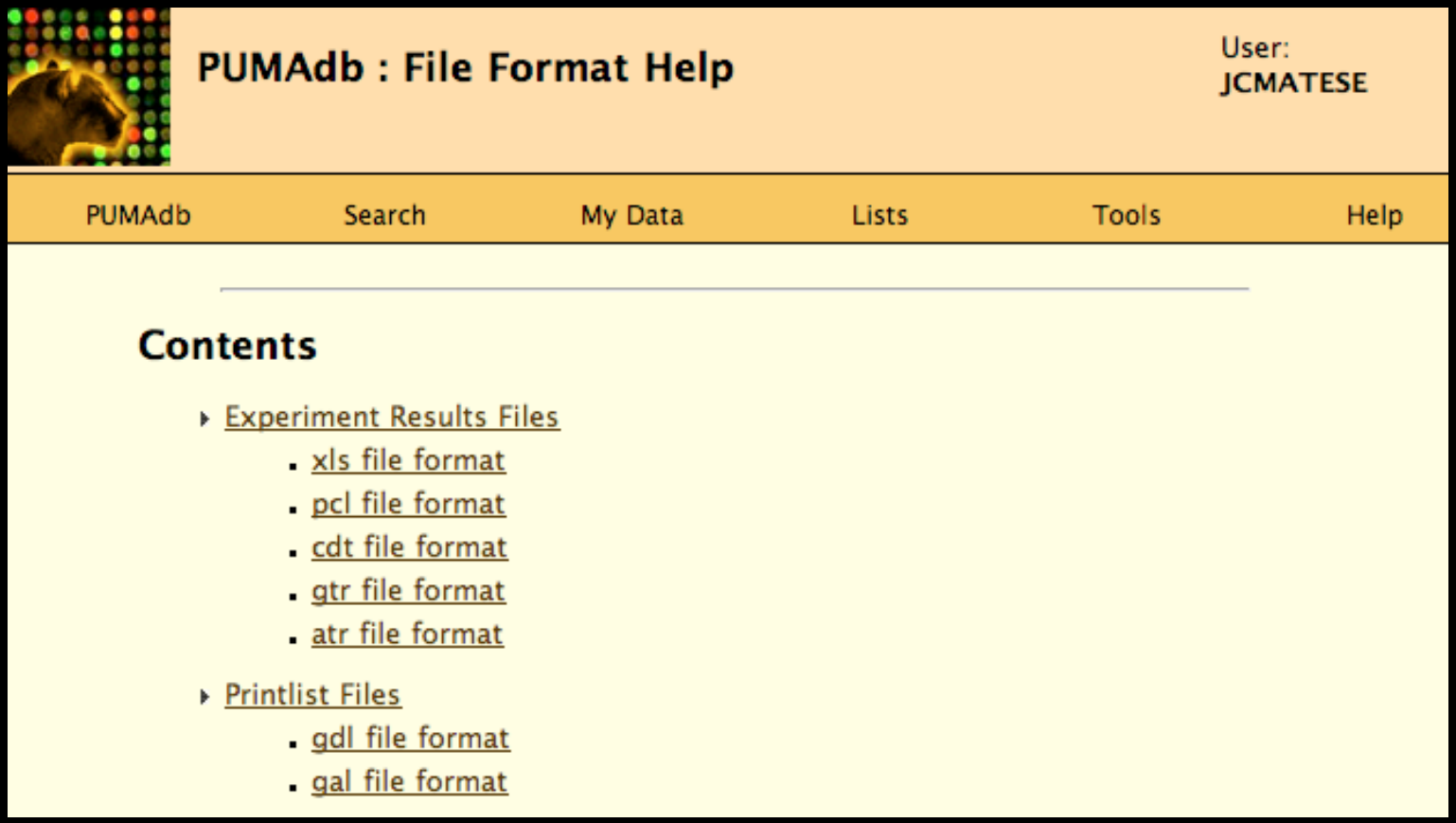
Data Repository Help



- How to use the tools
- Sharing data
- Options
- Links to help for
  - analysis methods,
  - data file formats
  - data retrieval and clustering

http://puma.princeton.edu/help/repository.shtml

File Format Help



http://puma.princeton.edu/help/formats.shtml

File Formats: Pre-clustering (PCL) File

UID is the Unique Identifier for the Spot/ Reporter

NAME sequence label for the Spot/Reporter

EWEIGHT indicates the weight the Array/ Assay is given in clustering

GWEIGHT indicates the weight the Spot/ Reporter is given in clustering

Names and orders of assays (if assays are not clustered)

Matrix values are for each spot/reporter on each array (usually log ratios)

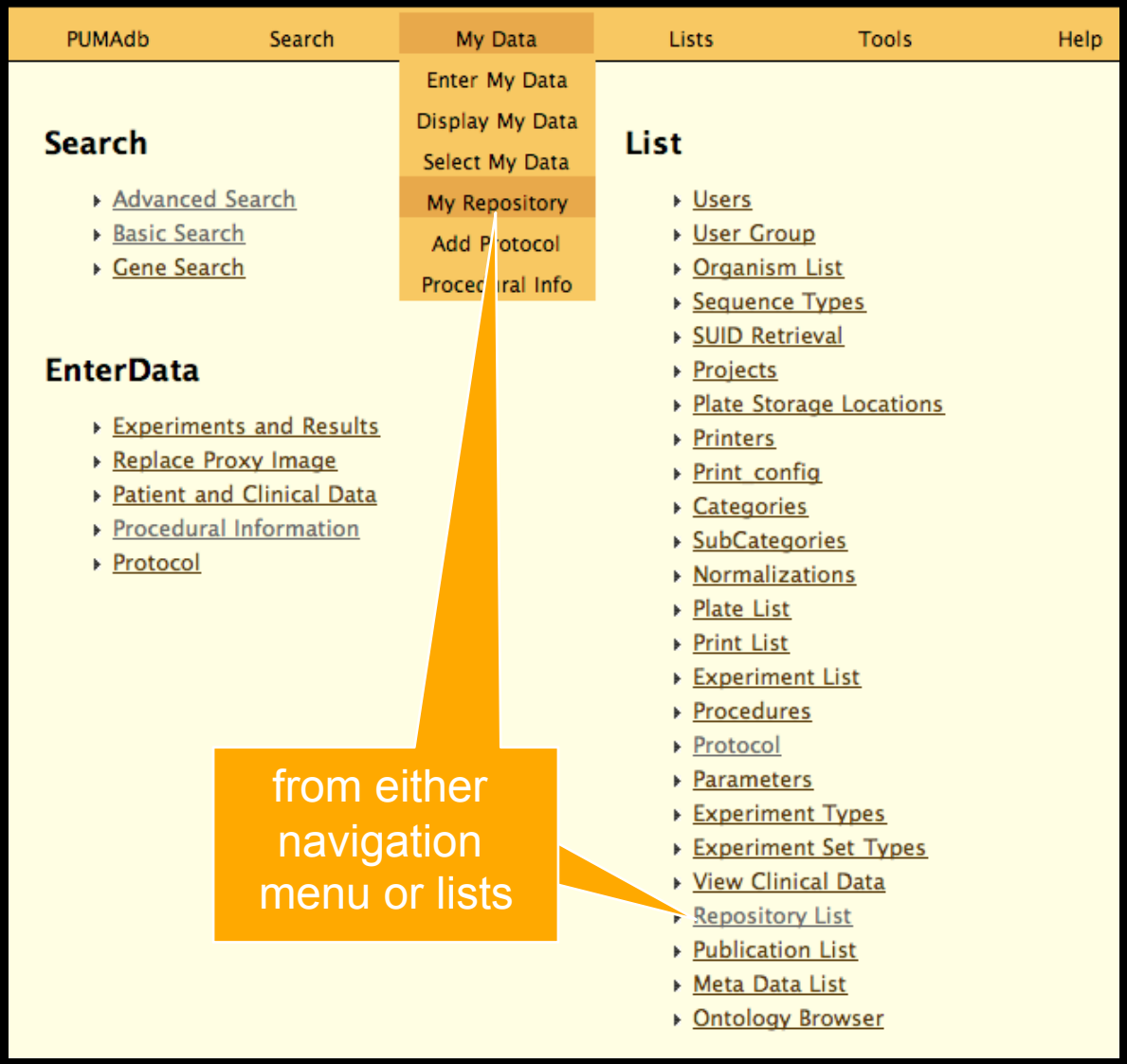
sample.txt

	A	B	C	D	E	F	G	H	I	J
1	UID	NAME	GWEIGHT	spo0	spo30	spo2	spo5	spo7	spo9	spo11
2	EWEIGHT									
3	[illegible]	EF		0.23	-1.79	-1.29	-1.56		-0.27	
				0.41	-0.38			-1.6	-1.84	-1.6
				0.61	-0.07			-2	-1.84	-2.25
				0.16	-0.15			-1.89	-1.74	-1.6
				0.03	1.39			-2.84	-2.47	-2.4
				-0.18	-0.18			-1.69	-1.43	-1.79
				-0.51	-0.62			4.54	3.22	4.33
				-0.14	-3.32			-2.4	-1.03	-0.6
11	YAL034C	FUN19	1	0.19	-0.03			-1.84	-1.94	-1.74
12	YAL035W	FUN12	1	0.01	-1.47	-1.15	-0.69	-1.36	-1.64	-1.29
13	YAL036C	FUN11	1	-0.15	-2.74	-1.79	-1.32	-2.12	0.3	-0.89
14	YAL038W	CDC19	1	-0.06	-1.89	-1.69	-2.32	-2.4	-0.81	-1.6
15	YAL040C	CLN3	1	-0.17	-2.25	-1.69	-2.25	-2.56	-0.3	-2.4
16	YAL054C	ACS1	1	0.51	2.6			1.35	-0.03	-0.23
17	YAL055W	YAL055W	1	-0.32	0.83				2.05	2.24
18	YAL062W	GDH3	1	0.3	2.59				0.34	1.36
19	YAL067C	SEO1	1	-0.17	3.44				1.61	2.8
20	YAR003W	YAR003W	1	-0.29	0.54				1.86	1.42
21	YAR007C	RFA1	1	-0.14	1.74				0.57	0.84
22	YAR015W	ADE1	1	0.11	-1.51				-1.89	-2
23	YAR027W	YAR027W	1	0.24	-1.06	-1.36		-1.25	-0.94	-1.36
24	YBL009W	YBL009W	1	-0.01	0.62	1.04	1.3	2.52	2.15	2.24
25	YBL010C	YBL010C	1	0.01	0.21	0.7	1.45	2.25	1.77	1.24
26	YBL015W	ACH1	1	0.52	1.01	1.49	1.75	1.49	0.58	0.19
27	YBL027W	RPL19A	1	0.01	-1.84	-0.97	-1.47	-1.79	-1	-0.51

What is the Repository?

- A method to save data sets to prevent repeatedly performing the same data retrieval
- A method to share processed data with others
- A way we can provide you with web access to new and/or computationally-intensive tools

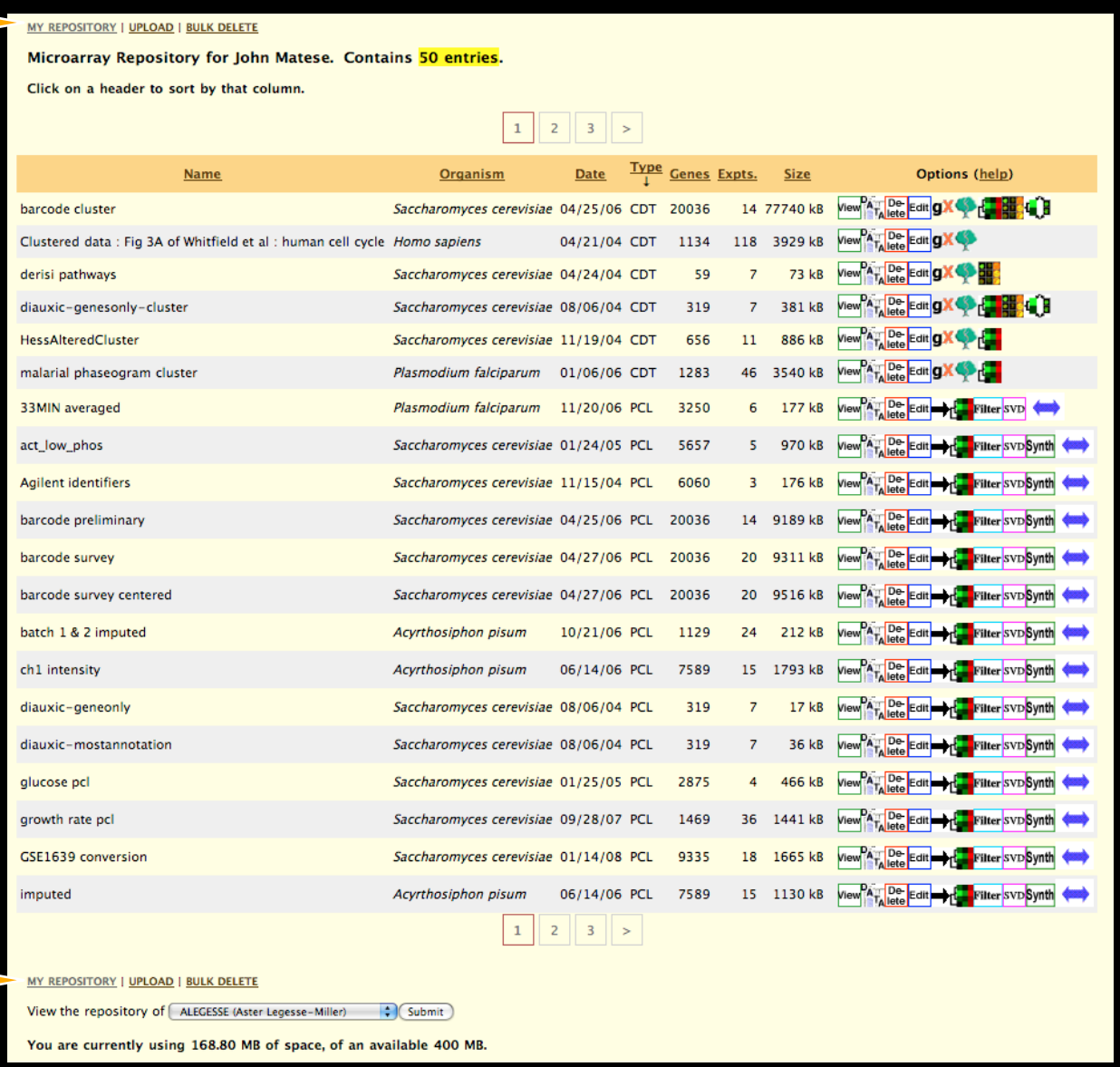
Accessing the Repository



Repository overview

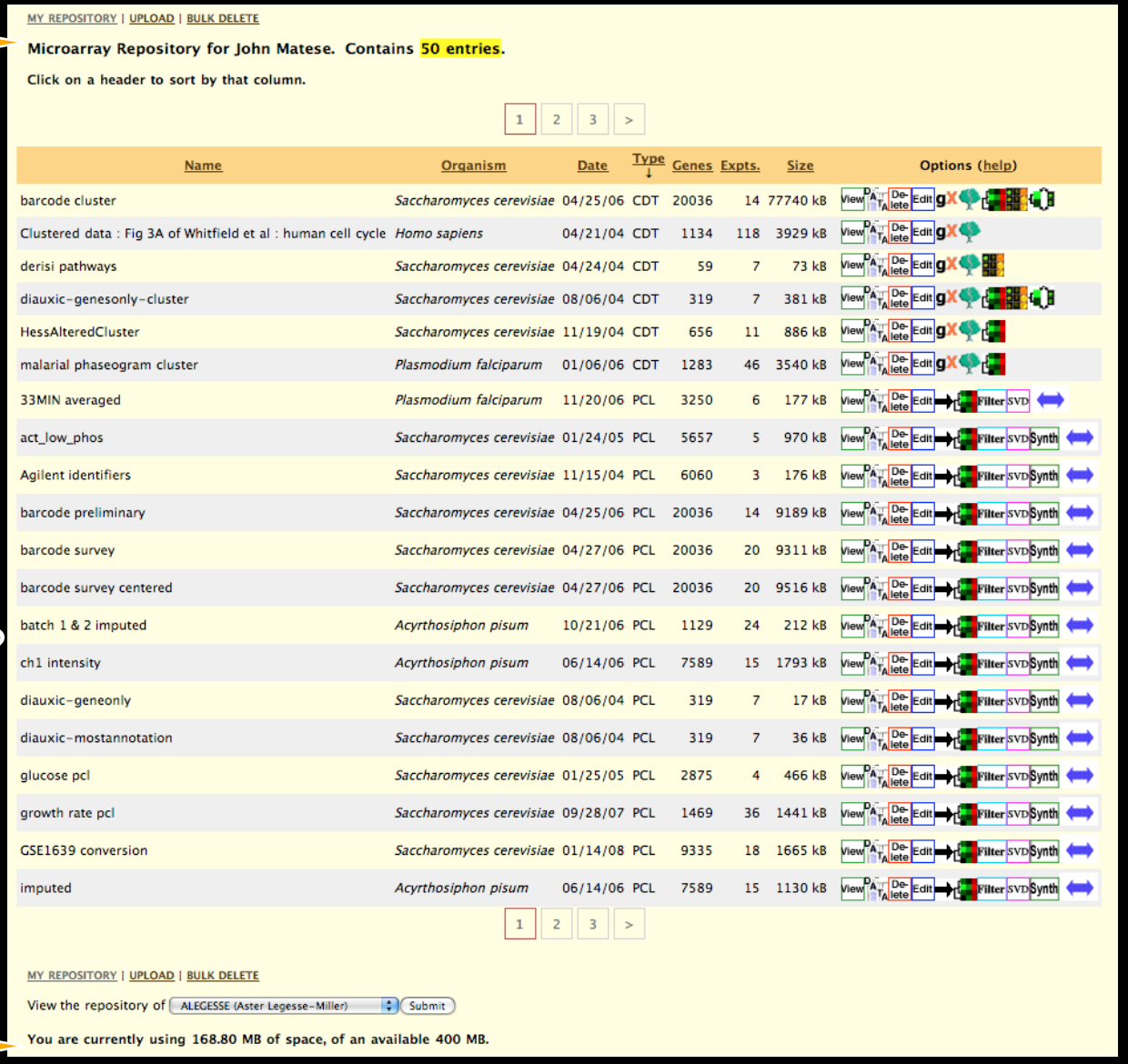
Links

- "Home",
- Upload,
- Delete



Repository overview

- Which repository?
- How many deposits?
- How much space?



MY REPOSITORY | UPLOAD | BULK DELETE

Microarray Repository for John Matese. Contains 50 entries.

Click on a header to sort by that column.

1 2 3 >

Name	Organism	Date	Type	Genes	Expts.	Size	Options (help)
barcode cluster	*Saccharomyces cerevisiae*	04/25/06	CDT	20036	14	77740 kB	View Data Delete Edit gX
Clustered data : Fig 3A of Whitfield et al : human cell cycle	*Homo sapiens*	04/21/04	CDT	1134	118	3929 kB	View Data Delete Edit gX
derisi pathways	*Saccharomyces cerevisiae*	04/24/04	CDT	59	7	73 kB	View Data Delete Edit gX
diauxic-genesonly-cluster	*Saccharomyces cerevisiae*	08/06/04	CDT	319	7	381 kB	View Data Delete Edit gX
HessAlteredCluster	*Saccharomyces cerevisiae*	11/19/04	CDT	656	11	886 kB	View Data Delete Edit gX
malarial phaseogram cluster	*Plasmodium falciparum*	01/06/06	CDT	1283	46	3540 kB	View Data Delete Edit gX
33MIN averaged	*Plasmodium falciparum*	11/20/06	PCL	3250	6	177 kB	View Data Delete Edit Filter SVD
act_low_phos	*Saccharomyces cerevisiae*	01/24/05	PCL	5657	5	970 kB	View Data Delete Edit Filter SVD Synth
Agilent identifiers	*Saccharomyces cerevisiae*	11/15/04	PCL	6060	3	176 kB	View Data Delete Edit Filter SVD Synth
barcode preliminary	*Saccharomyces cerevisiae*	04/25/06	PCL	20036	14	9189 kB	View Data Delete Edit Filter SVD Synth
barcode survey	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9311 kB	View Data Delete Edit Filter SVD Synth
barcode survey centered	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9516 kB	View Data Delete Edit Filter SVD Synth
batch 1 & 2 imputed	*Acyrthosiphon pisum*	10/21/06	PCL	1129	24	212 kB	View Data Delete Edit Filter SVD Synth
ch1 intensity	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1793 kB	View Data Delete Edit Filter SVD Synth
diauxic-geneonly	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	17 kB	View Data Delete Edit Filter SVD Synth
diauxic-mostannotation	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	36 kB	View Data Delete Edit Filter SVD Synth
glucose pcl	*Saccharomyces cerevisiae*	01/25/05	PCL	2875	4	466 kB	View Data Delete Edit Filter SVD Synth
growth rate pcl	*Saccharomyces cerevisiae*	09/28/07	PCL	1469	36	1441 kB	View Data Delete Edit Filter SVD Synth
GSE1639 conversion	*Saccharomyces cerevisiae*	01/14/08	PCL	9335	18	1665 kB	View Data Delete Edit Filter SVD Synth
imputed	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1130 kB	View Data Delete Edit Filter SVD Synth

1 2 3 >

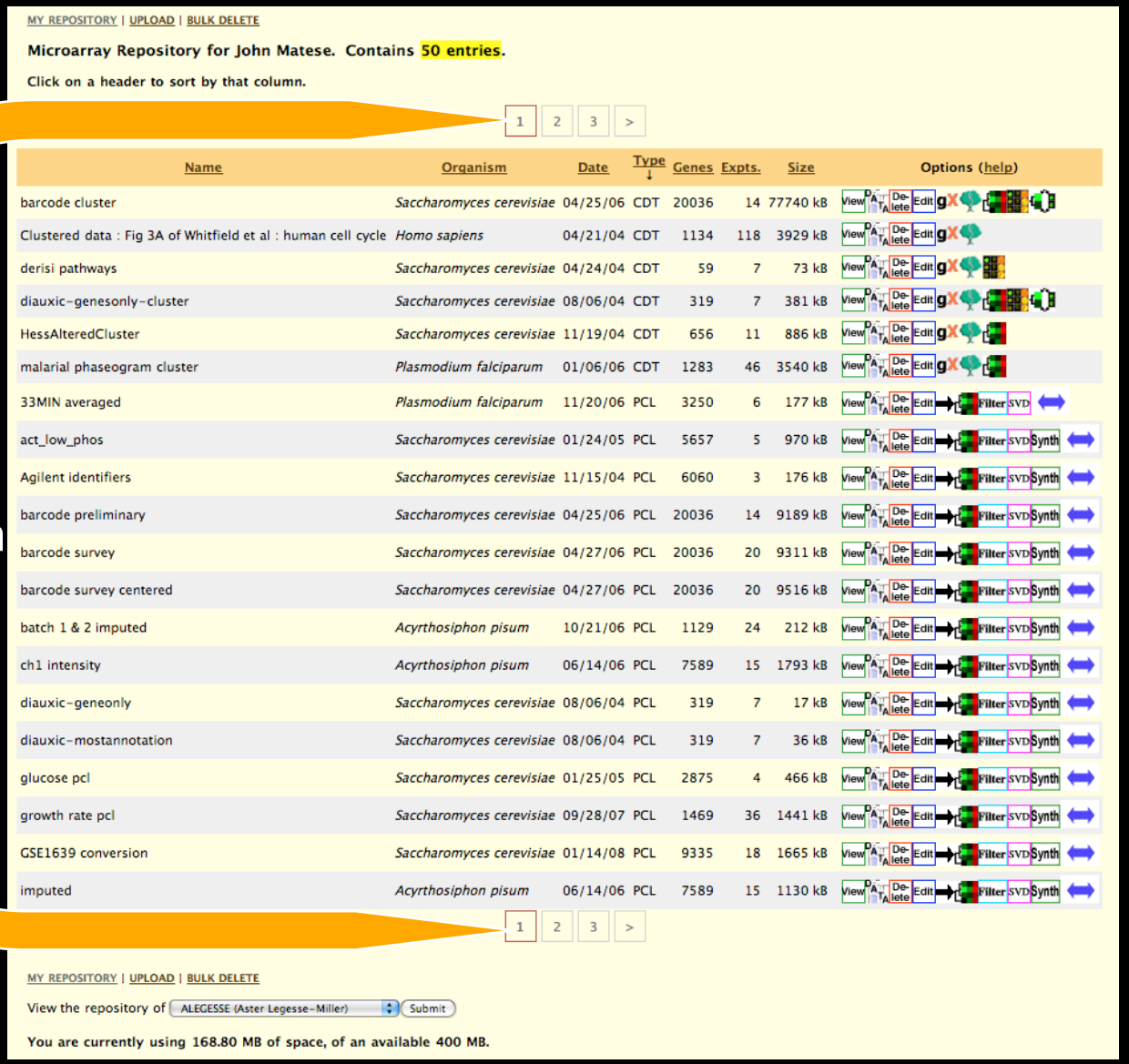
MY REPOSITORY | UPLOAD | BULK DELETE

View the repository of ALEGESSE (Aster Legesse-Miller) Submit

You are currently using 168.80 MB of space, of an available 400 MB.

Repository overview

Page Navigation



MY REPOSITORY | UPLOAD | BULK DELETE

Microarray Repository for John Matese. Contains 50 entries.

Click on a header to sort by that column.

1 2 3 >

Name	Organism	Date	Type	Genes	Expts.	Size	Options (help)
barcode cluster	*Saccharomyces cerevisiae*	04/25/06	CDT	20036	14	77740 kB	View Data Delete Edit gX
Clustered data : Fig 3A of Whitfield et al : human cell cycle	*Homo sapiens*	04/21/04	CDT	1134	118	3929 kB	View Data Delete Edit gX
derisi pathways	*Saccharomyces cerevisiae*	04/24/04	CDT	59	7	73 kB	View Data Delete Edit gX
diauxic-genesonly-cluster	*Saccharomyces cerevisiae*	08/06/04	CDT	319	7	381 kB	View Data Delete Edit gX
HessAlteredCluster	*Saccharomyces cerevisiae*	11/19/04	CDT	656	11	886 kB	View Data Delete Edit gX
malarial phaseogram cluster	*Plasmodium falciparum*	01/06/06	CDT	1283	46	3540 kB	View Data Delete Edit gX
33MIN averaged	*Plasmodium falciparum*	11/20/06	PCL	3250	6	177 kB	View Data Delete Edit Filter SVD
act_low_phos	*Saccharomyces cerevisiae*	01/24/05	PCL	5657	5	970 kB	View Data Delete Edit Filter SVD Synth
Agilent identifiers	*Saccharomyces cerevisiae*	11/15/04	PCL	6060	3	176 kB	View Data Delete Edit Filter SVD Synth
barcode preliminary	*Saccharomyces cerevisiae*	04/25/06	PCL	20036	14	9189 kB	View Data Delete Edit Filter SVD Synth
barcode survey	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9311 kB	View Data Delete Edit Filter SVD Synth
barcode survey centered	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9516 kB	View Data Delete Edit Filter SVD Synth
batch 1 & 2 imputed	*Acyrthosiphon pisum*	10/21/06	PCL	1129	24	212 kB	View Data Delete Edit Filter SVD Synth
ch1 intensity	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1793 kB	View Data Delete Edit Filter SVD Synth
diauxic-geneonly	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	17 kB	View Data Delete Edit Filter SVD Synth
diauxic-mostannotation	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	36 kB	View Data Delete Edit Filter SVD Synth
glucose pcl	*Saccharomyces cerevisiae*	01/25/05	PCL	2875	4	466 kB	View Data Delete Edit Filter SVD Synth
growth rate pcl	*Saccharomyces cerevisiae*	09/28/07	PCL	1469	36	1441 kB	View Data Delete Edit Filter SVD Synth
GSE1639 conversion	*Saccharomyces cerevisiae*	01/14/08	PCL	9335	18	1665 kB	View Data Delete Edit Filter SVD Synth
imputed	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1130 kB	View Data Delete Edit Filter SVD Synth

1 2 3 >

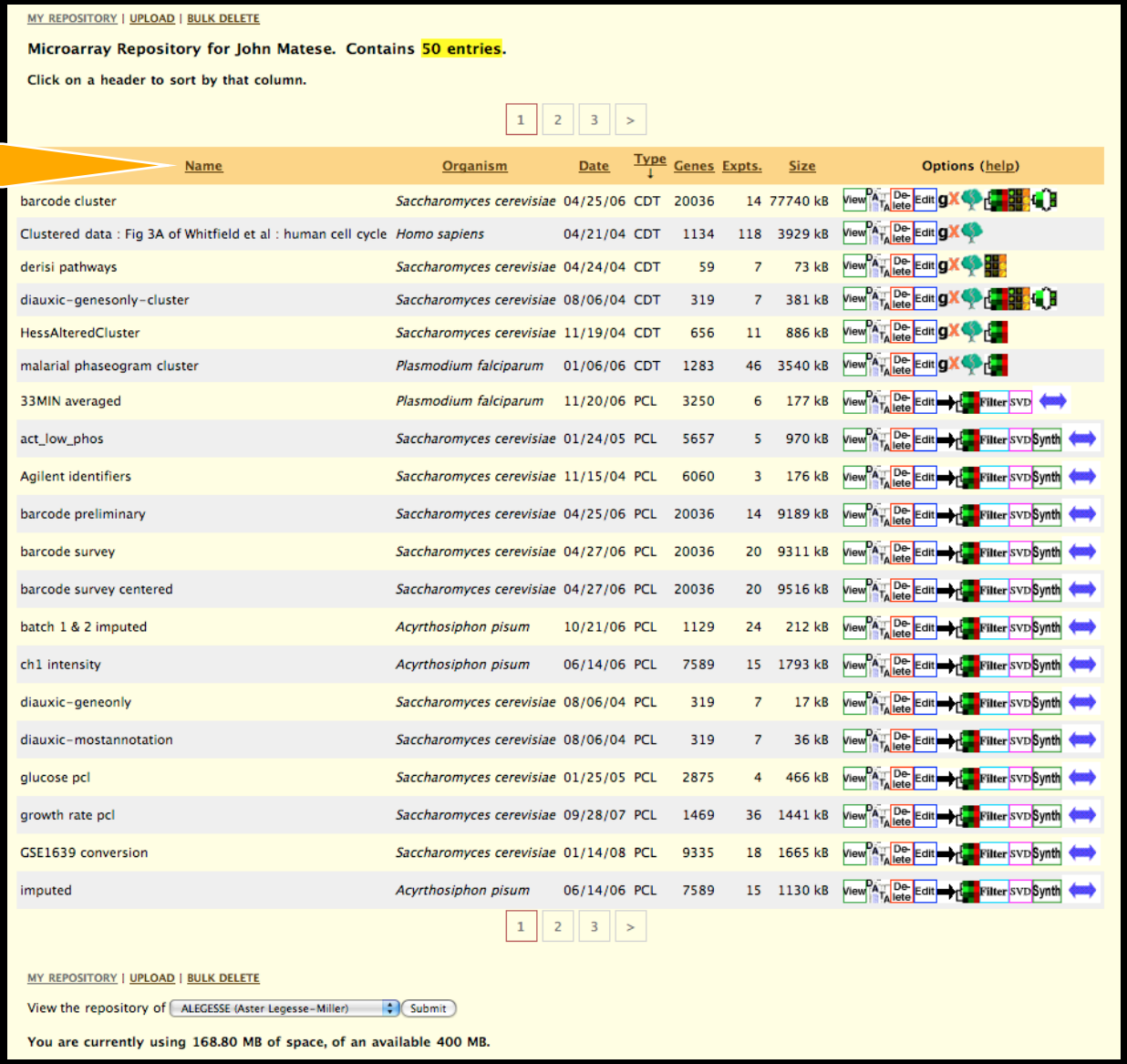
MY REPOSITORY | UPLOAD | BULK DELETE

View the repository of ALEGESSE (Aster Legesse-Miller) Submit

You are currently using 168.80 MB of space, of an available 400 MB.

Repository overview

Deposit
Sorting



MY REPOSITORY | UPLOAD | BULK DELETE

Microarray Repository for John Matese. Contains 50 entries.

Click on a header to sort by that column.

1 2 3 >

Name	Organism	Date	Type ↓	Genes	Expts.	Size	Options (help)
barcode cluster	*Saccharomyces cerevisiae*	04/25/06	CDT	20036	14	77740 kB	View Data Delete Edit gX
Clustered data : Fig 3A of Whitfield et al : human cell cycle	*Homo sapiens*	04/21/04	CDT	1134	118	3929 kB	View Data Delete Edit gX
derisi pathways	*Saccharomyces cerevisiae*	04/24/04	CDT	59	7	73 kB	View Data Delete Edit gX
diauxic-genesonly-cluster	*Saccharomyces cerevisiae*	08/06/04	CDT	319	7	381 kB	View Data Delete Edit gX
HessAlteredCluster	*Saccharomyces cerevisiae*	11/19/04	CDT	656	11	886 kB	View Data Delete Edit gX
malarial phaseogram cluster	*Plasmodium falciparum*	01/06/06	CDT	1283	46	3540 kB	View Data Delete Edit gX
33MIN averaged	*Plasmodium falciparum*	11/20/06	PCL	3250	6	177 kB	View Data Delete Edit Filter SVD
act_low_phos	*Saccharomyces cerevisiae*	01/24/05	PCL	5657	5	970 kB	View Data Delete Edit Filter SVD Synth
Agilent identifiers	*Saccharomyces cerevisiae*	11/15/04	PCL	6060	3	176 kB	View Data Delete Edit Filter SVD Synth
barcode preliminary	*Saccharomyces cerevisiae*	04/25/06	PCL	20036	14	9189 kB	View Data Delete Edit Filter SVD Synth
barcode survey	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9311 kB	View Data Delete Edit Filter SVD Synth
barcode survey centered	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9516 kB	View Data Delete Edit Filter SVD Synth
batch 1 & 2 imputed	*Acyrthosiphon pisum*	10/21/06	PCL	1129	24	212 kB	View Data Delete Edit Filter SVD Synth
ch1 intensity	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1793 kB	View Data Delete Edit Filter SVD Synth
diauxic-geneonly	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	17 kB	View Data Delete Edit Filter SVD Synth
diauxic-mostannotation	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	36 kB	View Data Delete Edit Filter SVD Synth
glucose pcl	*Saccharomyces cerevisiae*	01/25/05	PCL	2875	4	466 kB	View Data Delete Edit Filter SVD Synth
growth rate pcl	*Saccharomyces cerevisiae*	09/28/07	PCL	1469	36	1441 kB	View Data Delete Edit Filter SVD Synth
GSE1639 conversion	*Saccharomyces cerevisiae*	01/14/08	PCL	9335	18	1665 kB	View Data Delete Edit Filter SVD Synth
imputed	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1130 kB	View Data Delete Edit Filter SVD Synth

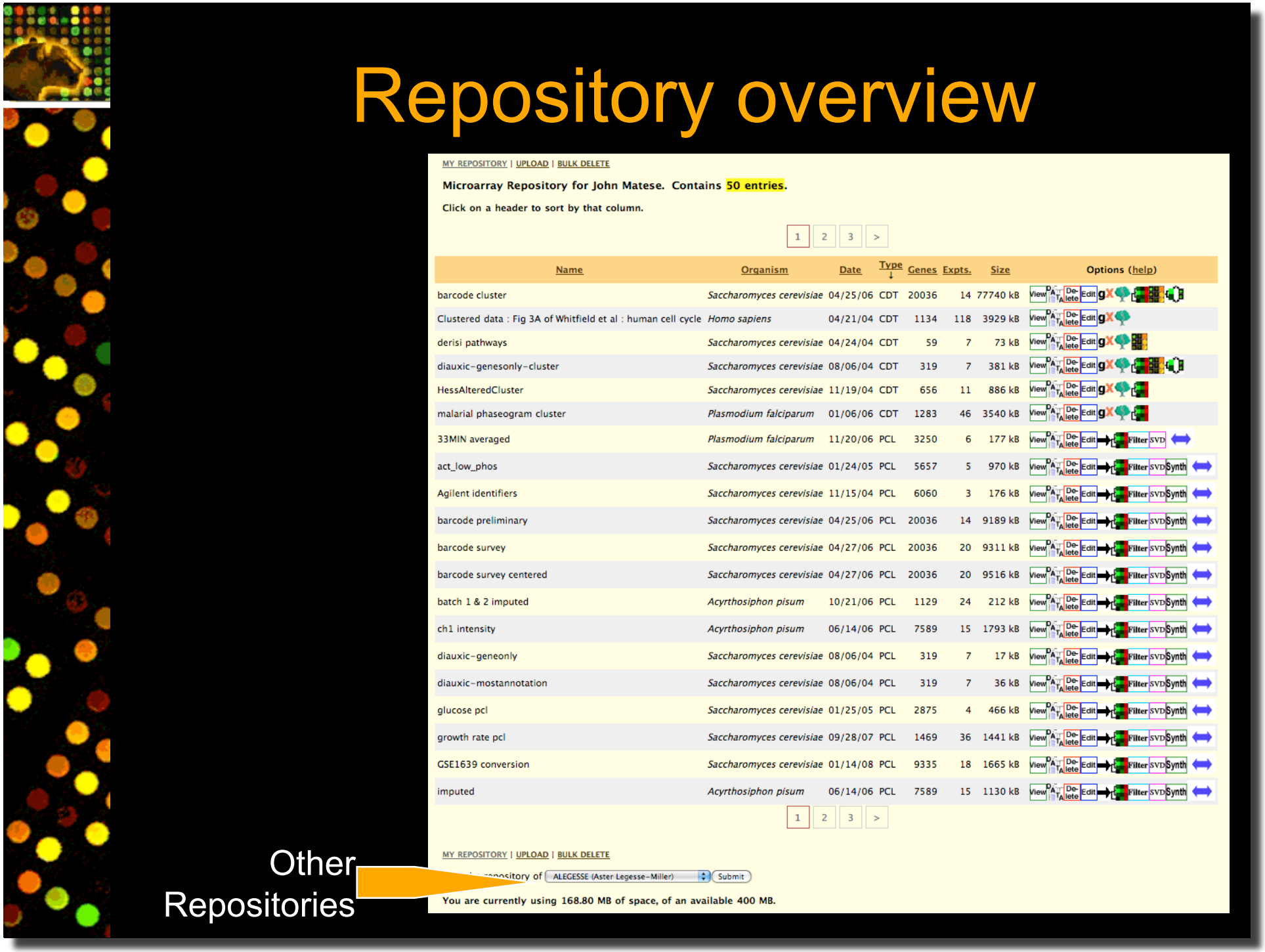
1 2 3 >

MY REPOSITORY | UPLOAD | BULK DELETE

View the repository of ALEGESSE (Aster Legesse-Miller) Submit

You are currently using 168.80 MB of space, of an available 400 MB.

Repository overview



MY REPOSITORY | UPLOAD | BULK DELETE

Microarray Repository for John Matese. Contains 50 entries.

Click on a header to sort by that column.

1 2 3 >

Name	Organism	Date	Type ↓	Genes	Expts.	Size	Options (help)
barcode cluster	*Saccharomyces cerevisiae*	04/25/06	CDT	20036	14	77740 kB	View Data Delete Edit gX
Clustered data : Fig 3A of Whitfield et al : human cell cycle	*Homo sapiens*	04/21/04	CDT	1134	118	3929 kB	View Data Delete Edit gX
derisi pathways	*Saccharomyces cerevisiae*	04/24/04	CDT	59	7	73 kB	View Data Delete Edit gX
diauxic-genesonly-cluster	*Saccharomyces cerevisiae*	08/06/04	CDT	319	7	381 kB	View Data Delete Edit gX
HessAlteredCluster	*Saccharomyces cerevisiae*	11/19/04	CDT	656	11	886 kB	View Data Delete Edit gX
malarial phaseogram cluster	*Plasmodium falciparum*	01/06/06	CDT	1283	46	3540 kB	View Data Delete Edit gX
33MIN averaged	*Plasmodium falciparum*	11/20/06	PCL	3250	6	177 kB	View Data Delete Edit Filter SVD
act_low_phos	*Saccharomyces cerevisiae*	01/24/05	PCL	5657	5	970 kB	View Data Delete Edit Filter SVD Synth
Agilent identifiers	*Saccharomyces cerevisiae*	11/15/04	PCL	6060	3	176 kB	View Data Delete Edit Filter SVD Synth
barcode preliminary	*Saccharomyces cerevisiae*	04/25/06	PCL	20036	14	9189 kB	View Data Delete Edit Filter SVD Synth
barcode survey	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9311 kB	View Data Delete Edit Filter SVD Synth
barcode survey centered	*Saccharomyces cerevisiae*	04/27/06	PCL	20036	20	9516 kB	View Data Delete Edit Filter SVD Synth
batch 1 & 2 imputed	*Acyrthosiphon pisum*	10/21/06	PCL	1129	24	212 kB	View Data Delete Edit Filter SVD Synth
ch1 intensity	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1793 kB	View Data Delete Edit Filter SVD Synth
diauxic-geneonly	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	17 kB	View Data Delete Edit Filter SVD Synth
diauxic-mostannotation	*Saccharomyces cerevisiae*	08/06/04	PCL	319	7	36 kB	View Data Delete Edit Filter SVD Synth
glucose pcl	*Saccharomyces cerevisiae*	01/25/05	PCL	2875	4	466 kB	View Data Delete Edit Filter SVD Synth
growth rate pcl	*Saccharomyces cerevisiae*	09/28/07	PCL	1469	36	1441 kB	View Data Delete Edit Filter SVD Synth
GSE1639 conversion	*Saccharomyces cerevisiae*	01/14/08	PCL	9335	18	1665 kB	View Data Delete Edit Filter SVD Synth
imputed	*Acyrthosiphon pisum*	06/14/06	PCL	7589	15	1130 kB	View Data Delete Edit Filter SVD Synth

1 2 3 >

MY REPOSITORY | UPLOAD | BULK DELETE

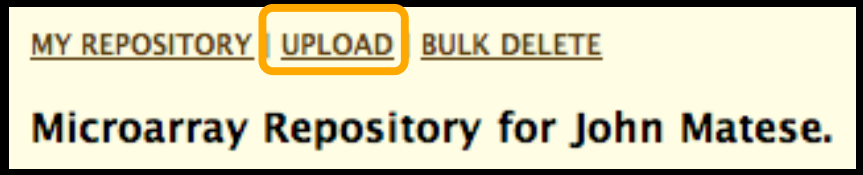
repository of ALEGESSE (Aster Legesse-Miller) Submit

You are currently using 168.80 MB of space, of an available 400 MB.

Other Repositories

Depositing Data into your Respository

- Deposit, from data retrieval “pipeline”



- Upload, from local workstation within repository itself

Repository Deposits : CDT

Name	Organism	Date	Type ↓	Genes	Expts.	Size	Options (help)
all AA and growth rate expts	*Saccharomyces cerevisiae*	02/05/07	CDT	5595	122	26345 kB	View
BDY expression	*Saccharomyces cerevisiae*	04/27/08	CDT	1883	11	5383 kB	View
BDYs	*Saccharomyces cerevisiae*	03/26/08	CDT	2395	15	7139 kB	View
ditri ordered clustered	*Saccharomyces cerevisiae*	11/12/07	CDT	2639	25	6149 kB	View

- View cluster using GeneXplorer or Treeview
- View cluster images
- View retrieval and clustering report
- Assign access

CDT Deposit Options

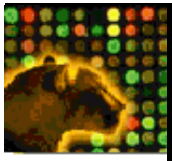
- 

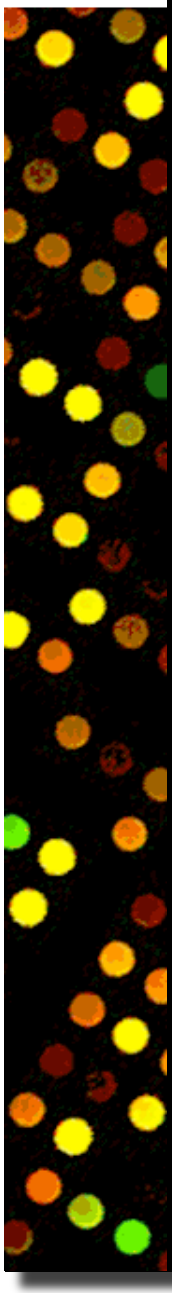
View retrieval report
- 
Download datafiles
- 

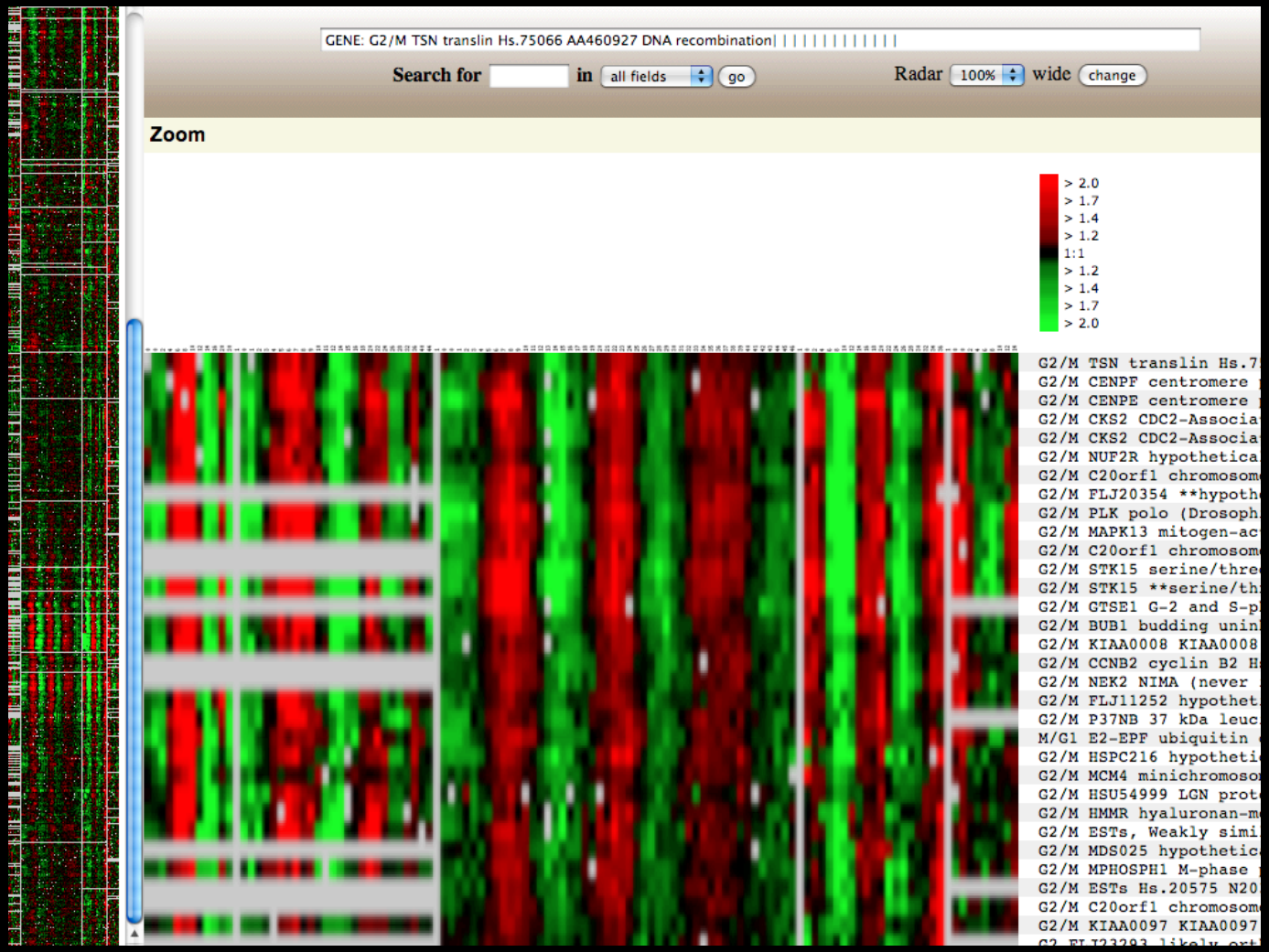
GeneXplorer (web browser)
- 
Java Treeview (web applet)
- 
Cluster Images (from pipeline)
- 
Clustered Spot Images (from pipeline)
- 
Combo Images (from pipeline)
- 

Edit deposit name, description, or access
- 

Delete deposit

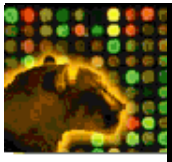


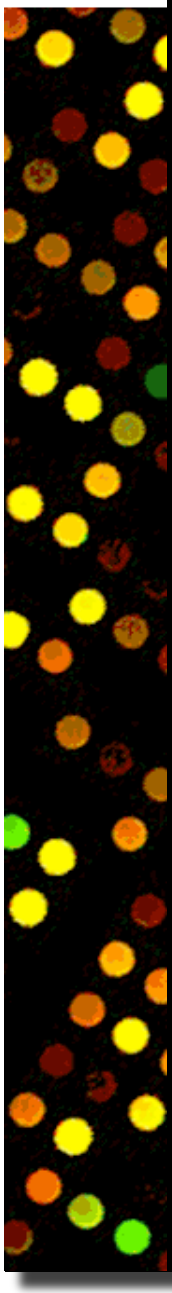


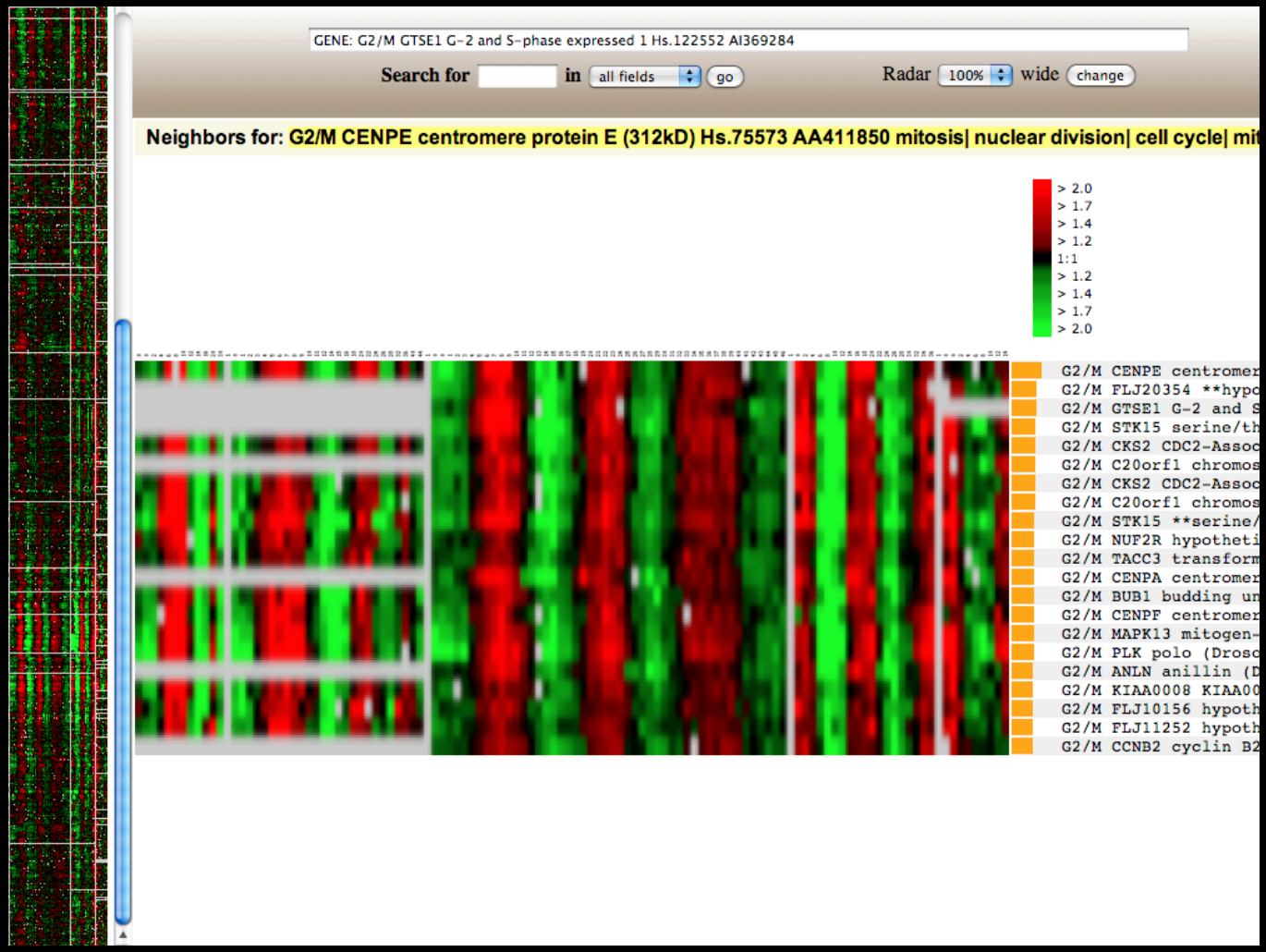


GeneXplorer : default views

- Radar
- Zoom
- Toolbar

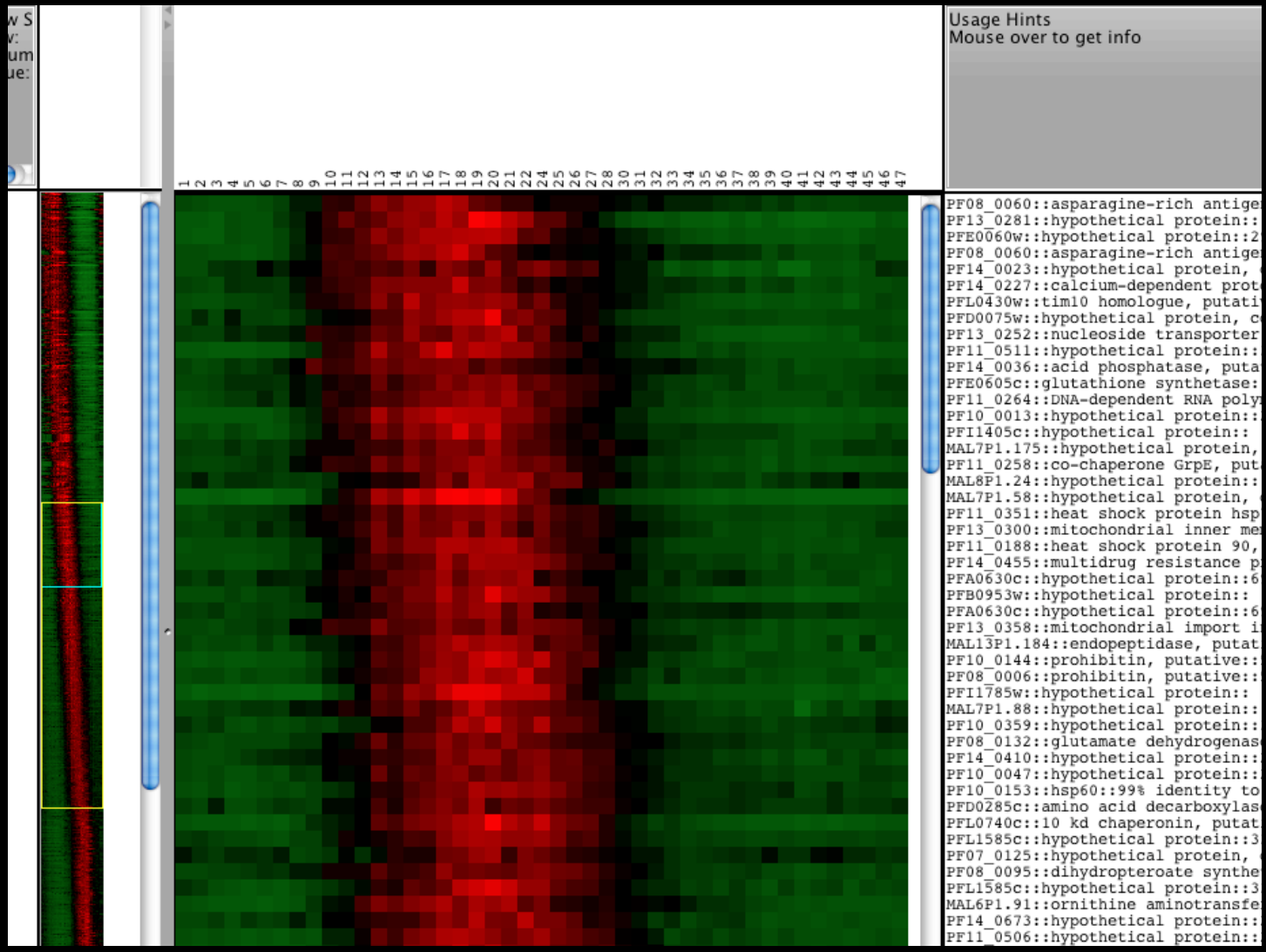






GeneXplorer : correlation view

Clicking on a vector within the zoom view display a list of vectors having the best correlations.



Java TreeView Web Applet

Provides a quick visualization of the dataset.

Repository Deposits : PCL

Name	Organism	Date	Type ↓	Genes	Expts.	Size	Options (help)
Ty red	*Saccharomyces cerevisiae*	06/06/07	PCL	41775	13	6783 kB	View Filter SVD Synth
YZRNADNA both median centered	*Saccharomyces cerevisiae*	05/04/08	PCL	4348	8	1234 kB	View Filter SVD Synth
YZRNADNA median centered arrays	*Saccharomyces cerevisiae*	05/04/08	PCL	4348	8	1199 kB	View Filter SVD Synth
all except first 2 Sb processed and clustered	*Saccharomyces bayanus*	01/02/08	PCL_UPLOAD	4840	169	4239 kB	View Filter SVD
HO data from Abram	*Saccharomyces cerevisiae*	02/11/08	PCL_UPLOAD	7926	12	1089 kB	View Filter SVD Synth
May1 RNADNA orthologs	*Saccharomyces cerevisiae*	05/01/08	PCL_UPLOAD	6277	56	3120 kB	View Filter SVD Synth

- Deposit the results of an analysis
- Re-enter data retrieval “pipeline” (filter, cluster), potentially avoiding data retrieval procedures that can be both repetitive and tedious!
- Use Singular Value Decomposition (SVD) tools
- Average data by “synthetic genes”
- View retrieval report
- Assign access

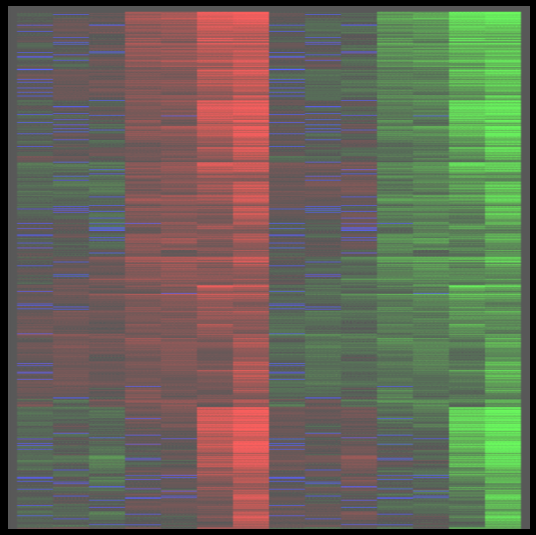
PCL Deposit Options

- View retrieval report
- Download datafiles
- Filter by gene expression pattern (pipeline re-entry)
- Cluster and visualize (pipeline re-entry)
- Singular value decomposition (SVD) tools
- Average data by “synthetic genes”
- Impute missing data
- Edit deposit name, description, or access
- Delete deposit

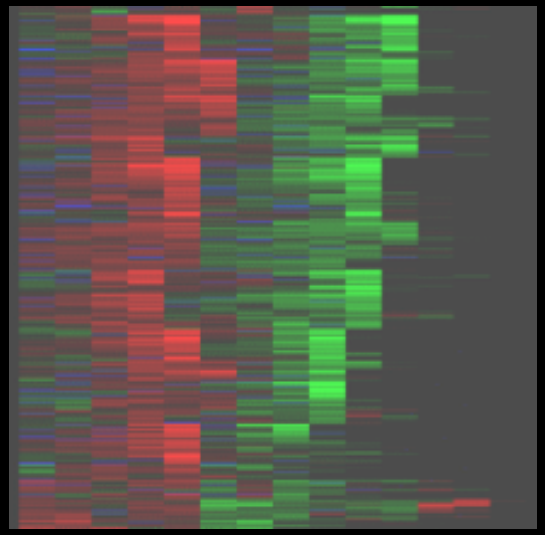
KNNImpute: The Missing Values Problem

- Microarrays can have systematic or random missing values
- Some algorithms aren't robust to missing values
- Large literature on parameter estimation exists
- What's best to do for microarrays?

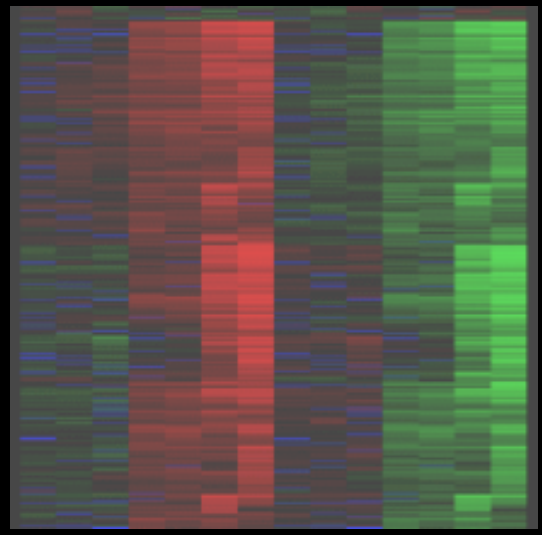
Why Estimate Missing Values?



Complete data set



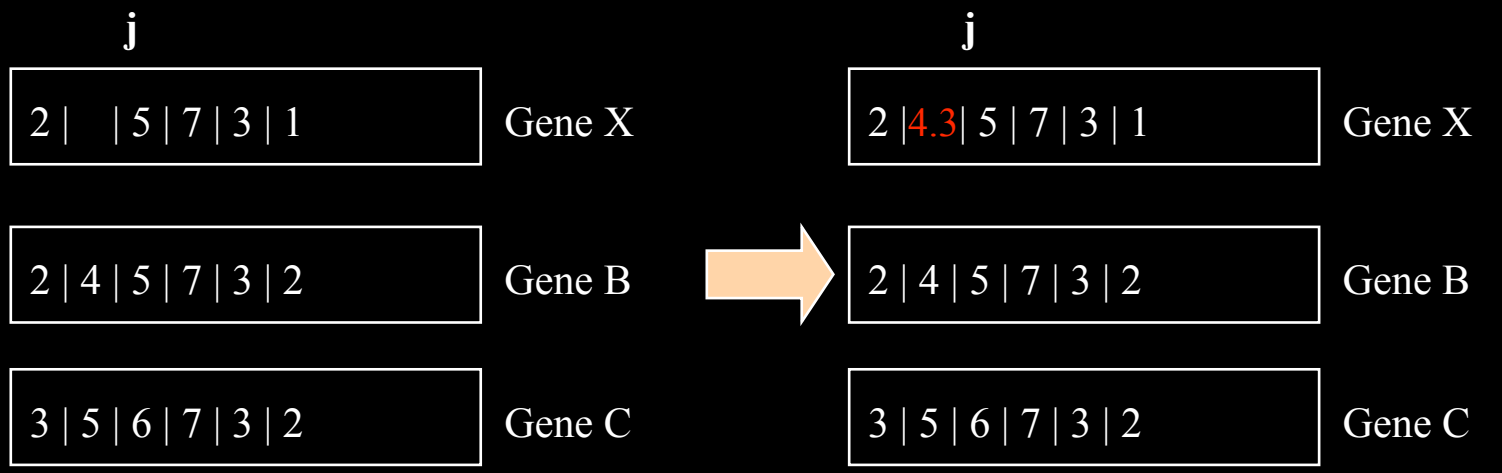
Data set with 30% entries missing (missing values appear black)



Data set with missing values estimated by KNNimpute algorithm

KNNimpute Algorithm

- Idea: use genes with similar expression profiles to estimate missing values



Troyanskaya *et al*. Missing value estimation methods for DNA microarrays. Bioinformatics (2001) vol. 17 (6) pp. 520-5

KNNImpute implementation

Background

Because some data analysis algorithms either require (i.e. SVD) or are sensitive to missing data, data imputation is sometimes necessary. The interface provided here initiates an implementation of K-nearest neighbors data imputation (KNNimpute), as described in Troyanskaya et al. Bioinformatics 17 (6): 520. (2001) , and places it within the job queue. The resulting imputed dataset will be deposited to your repository.

Impute Parameters	
Number of nearest neighbors :	15
Distance metric :	euclidean *note: only euclidean has been implemented*
Resulting Deposit Name :	
Resulting Deposit Description:	describe the new imputed pcl file

Submit Impute Reset

Initial Dataset Characterization

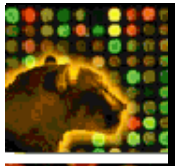
- What is the structure of the data?
- What are the patterns within the noise?
- Reduce a dataset's dimensionality, but capturing its variance
  - Principle component analysis (**PCA**)
  - Singular Value Decomposition (**SVD**)

http://public.lanl.gov/mewall/kluwer2002.html

"Singular value decomposition and principal component analysis" in ***A Practical Approach to Microarray Data Analysis (D.P. Berrar, W. Dubitzky, M. Granzow, eds.) Kluwer: Norwell, MA, 2003. pp. 91-109. LANL LA-UR-02-4001***

Reducing dimensionality

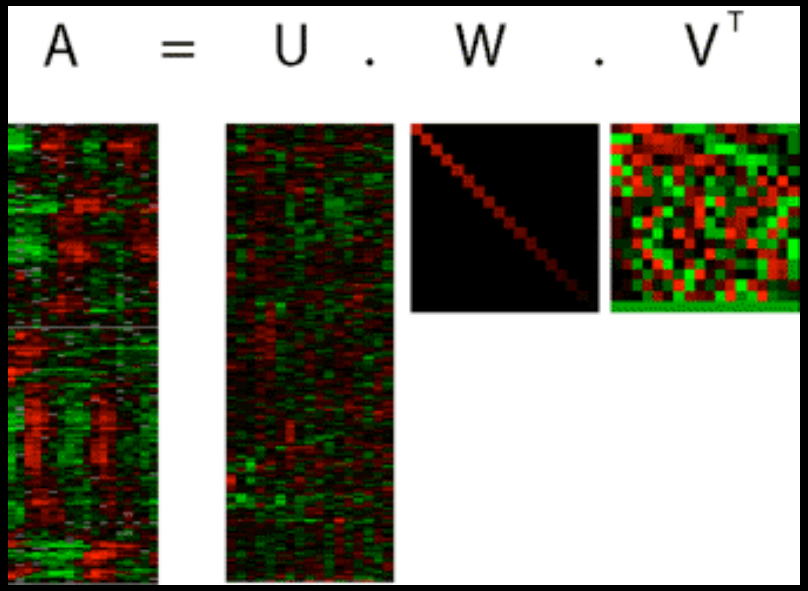
- Let's imagine we have a three-dimensional cigar, as shown in **A**
- We can represent this in one dimension, by looking at its lengthwise shadow (**B**)
- Looking at its cross-wise shadow (**C**), we get an orthogonal view of the cigar that tells us more about the three-dimensional object than **B** alone.



SVD: Singular Value Decomposition

- SVD determines a set of patterns that describe the greatest amount of variance in a dataset
- SVD determines unique orthogonal (or uncorrelated) gene and corresponding array expression patterns ("eigengenes" and "eigenarrays," respectively)
- These patterns often are correlated with…
  - biological processes
  - technical artifacts

SVD : Algorithm



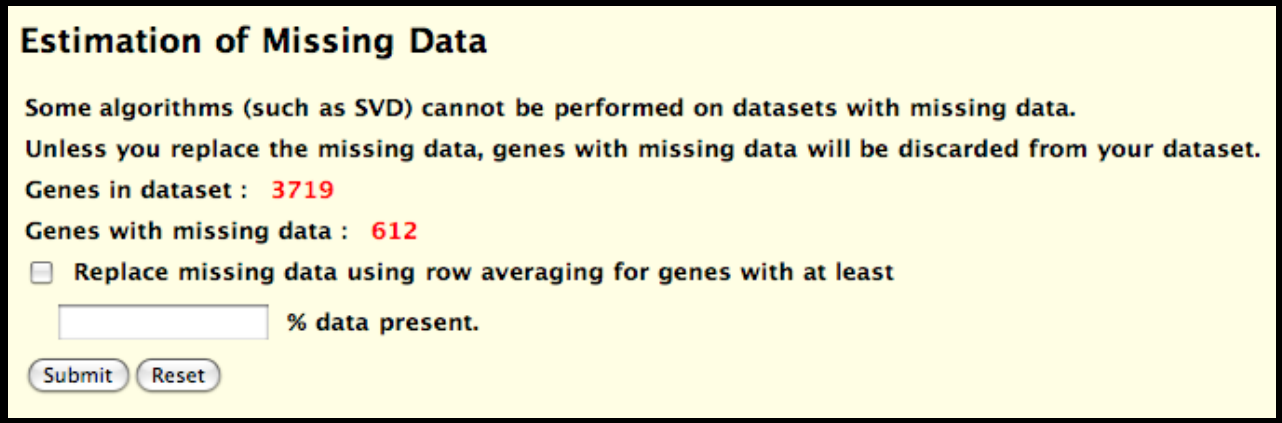
Singular value decomposition of a matrix of gene expression values (matrix **A) results in three matrices, U (the eigenarrays matrix), W (the eigenexpression levels matrix) and V^T (the eigengenes matrix). Matrix V^T contains the eigengenes, matrix W contains the eigenvalues, matrix U contains the coefficients for the genes for each eigengene.**

SVD is a linear transformation that uses a theorem from linear algebra to decompose a matrix to the product of three other matrices. For microarray data, the mathematical variables of SVD are the "eigengenes," "eigenarrays" and "eigenexpression levels." The eigengenes are gene patterns of expression. The eigenarrays are array patterns of expression. And, the eigenexpression levels indicate the dominance, and therefore also the significance, of these gene and array patterns of expression in the data. These matrices are shown. For more detail please see:

http://puma.princeton.edu/help/svd.shtml

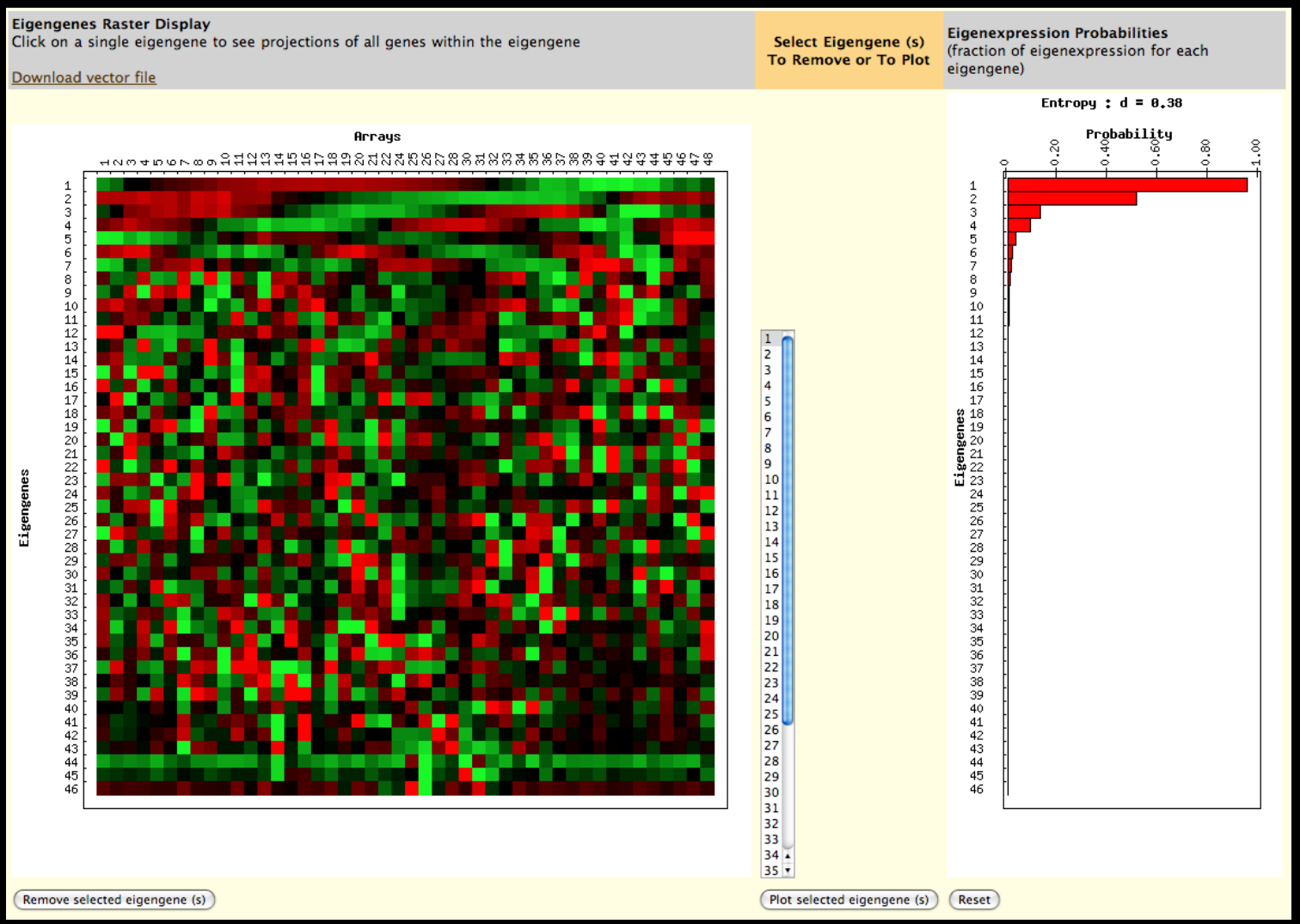
Alter *et al*. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA (2000) vol. 97 (18) pp. 10101-6

SVD: Missing Data Estimation

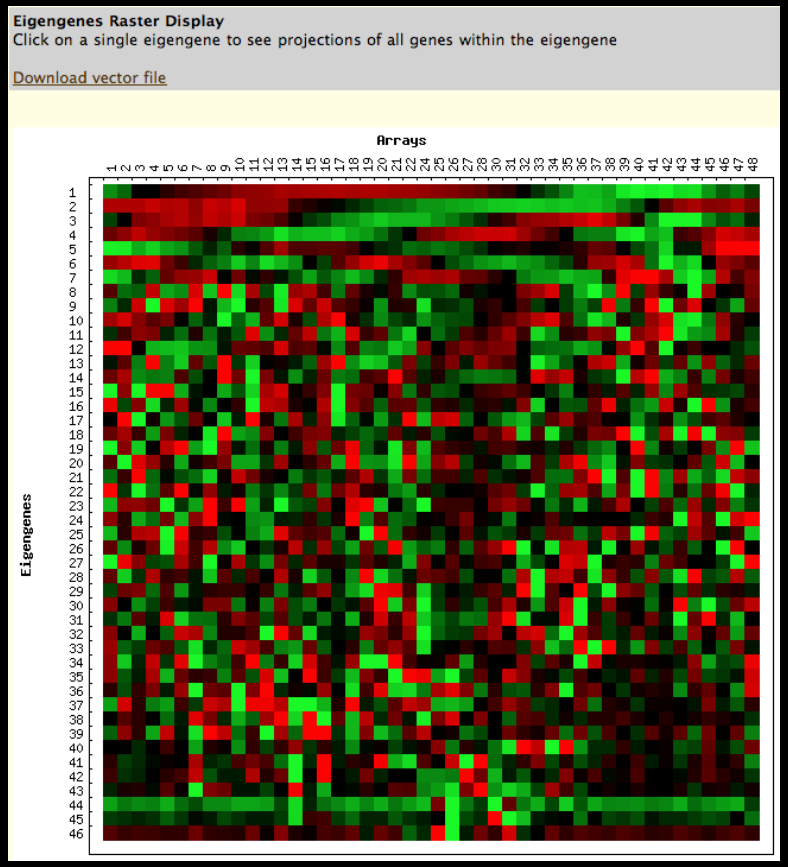


- Some algorithms (such as SVD) cannot operate with missing data
- You can use this simple method or you can use KNNImpute to estimate missing data

SVD Display

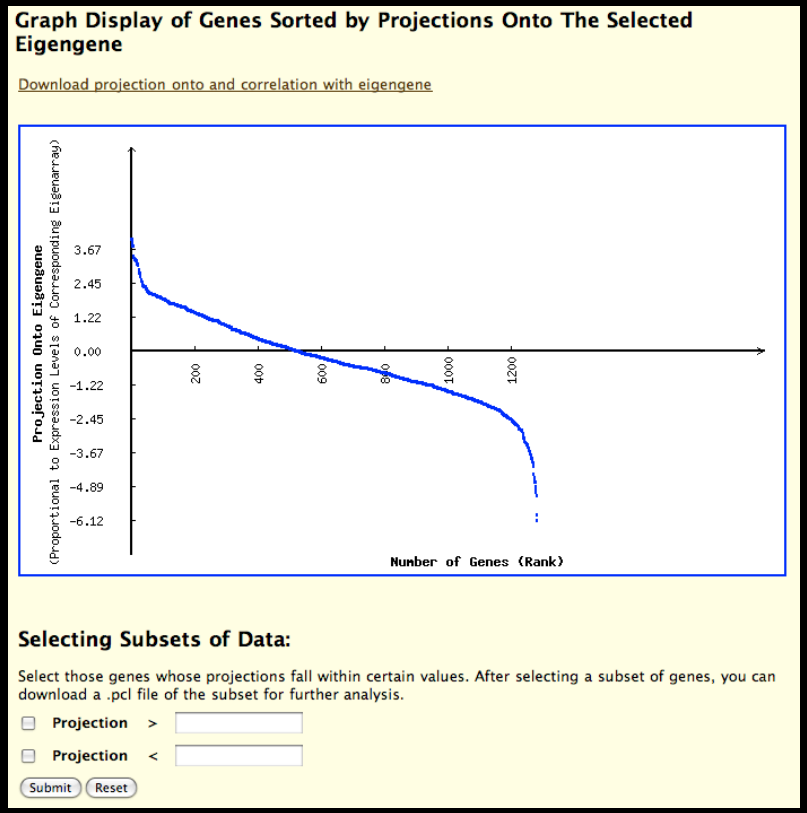

Eigengenes Raster Display
Click on a single eigengene to see projections of all genes within the eigengene
Download vector file
Select Eigengene (s)
To Remove or To Plot
Eigenexpression Probabilities
(fraction of eigenexpression for each
eigengene)
Arrays
Eigengenes
Entropy : d = 0.38
Probability
Remove selected eigengene (s)
Plot selected eigengene (s)
Reset

SVD: Raster Display



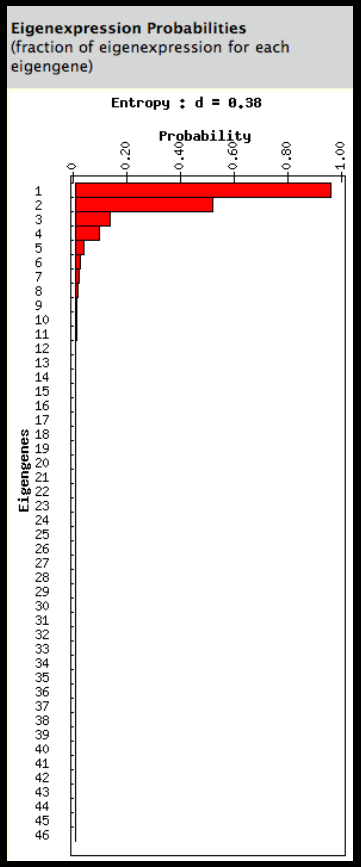
- Each row represents an “eigengene” -- an orthogonal representation of the genes in the dataset
- The topmost eigengene contributes the most to the data set

SVD: View Projection



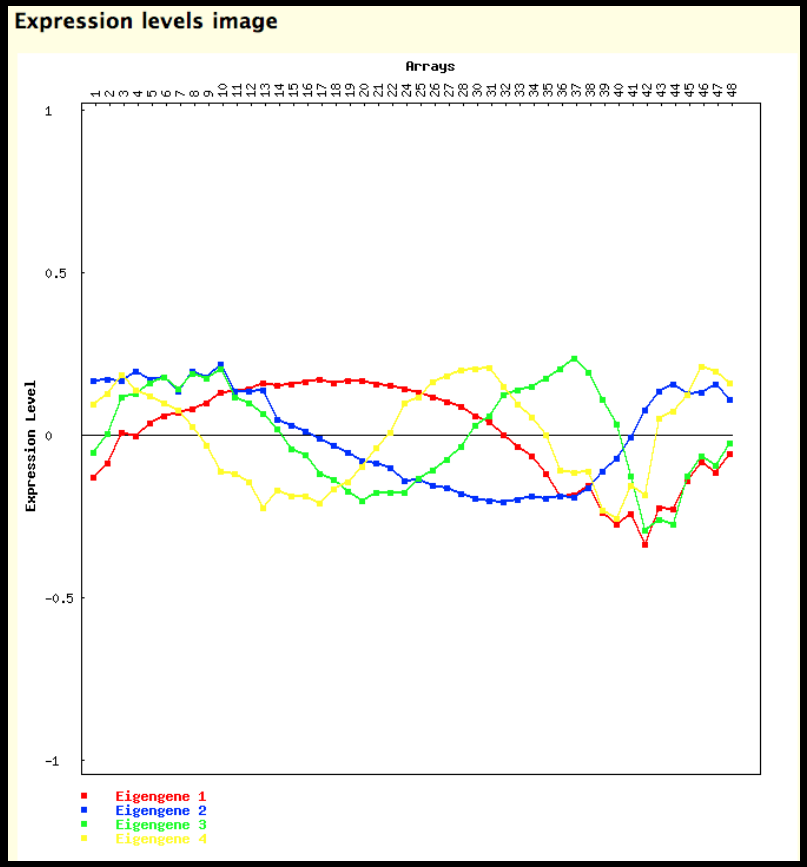
- Clicking on a row in the Raster Display brings you the Projection View
- You can select genes that have high and low contributions from an eigengene and download them in a PCL file
- In this way, you might use SVD to help classify subtypes

SVD: Eigenexpression



- Each bar show the probability of expression of each eigengene
- You can compare the probabilities to see which eigengenes contribute more to the overall “view” of the data

SVD: Plot selected eigengenes



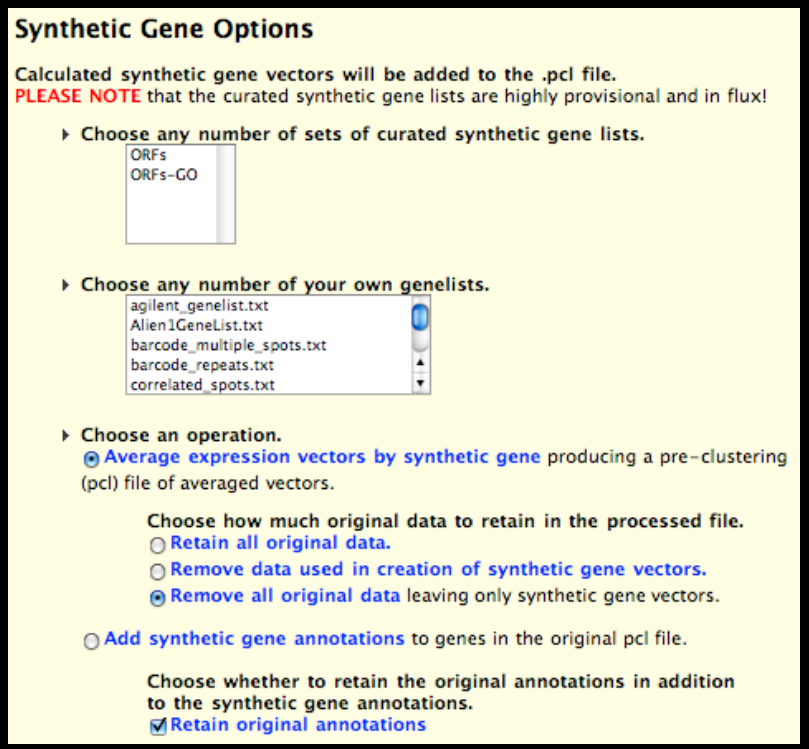
- You can plot as many or as few eigengenes as you like
- This plot gives you an easy-to-understand view of the behavior of each eigengene

Access to Synthetic Gene Tools

Name	Organism	Date	Type ↓	Genes	Expts.	Size	Options (help)
Ty red	*Saccharomyces cerevisiae*	06/06/07	PCL	41775	13	6783 kB	View Filter SVD Synth
YZRNADNA both median centered	*Saccharomyces cerevisiae*	05/04/08	PCL	4348	8	1234 kB	View Filter SVD Synth
YZRNADNA median centered arrays	*Saccharomyces cerevisiae*	05/04/08	PCL	4348	8	1199 kB	View Filter SVD Synth
all except first 2 Sb processed and clustered	*Saccharomyces bayanus*	01/02/08	PCL_UPLOAD	4840	169	4239 kB	View Filter SVD
HO data from Abram	*Saccharomyces cerevisiae*	02/11/08	PCL_UPLOAD	7926	12	1089 kB	View Filter SVD Synth
May1 RNADNA orthologs	*Saccharomyces cerevisiae*	05/01/08	PCL_UPLOAD	6277	56	3120 kB	View Filter SVD Synth

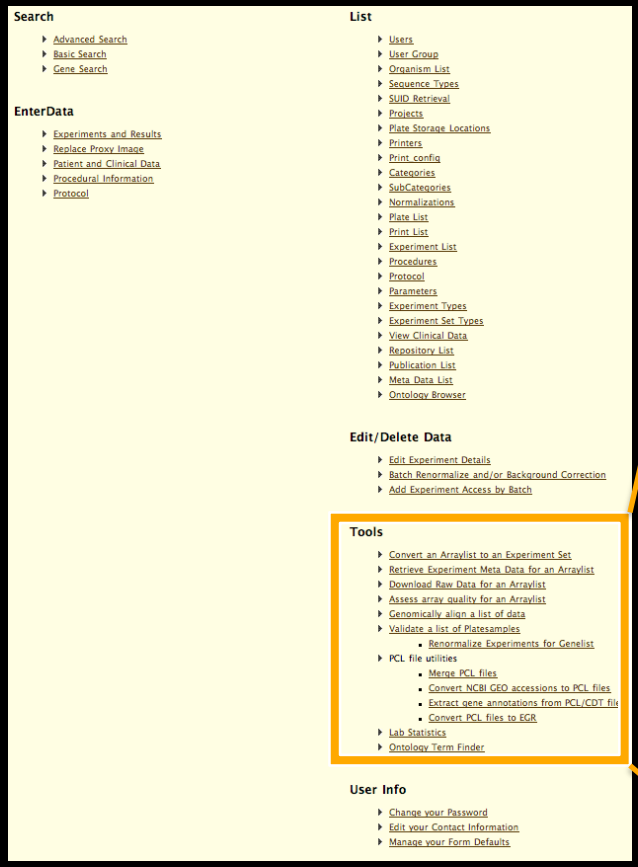
- Put data in your repository.
- Click “Synth” icon.

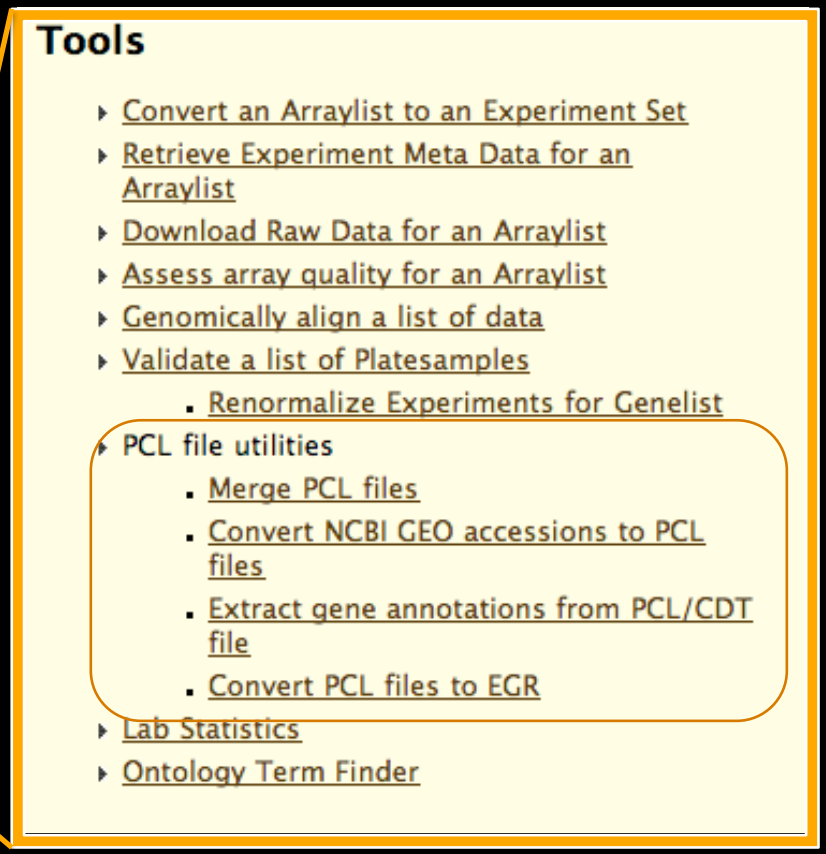
Synthetic Gene Options



- Select any number of public lists, and/or your own genelists.
- Choose handling of original data

Other PCL tools

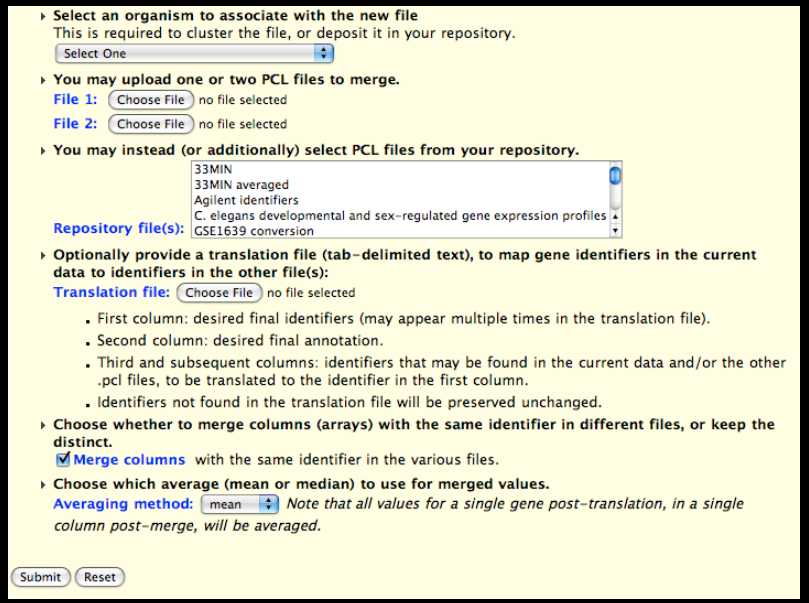




Merge PCL Files

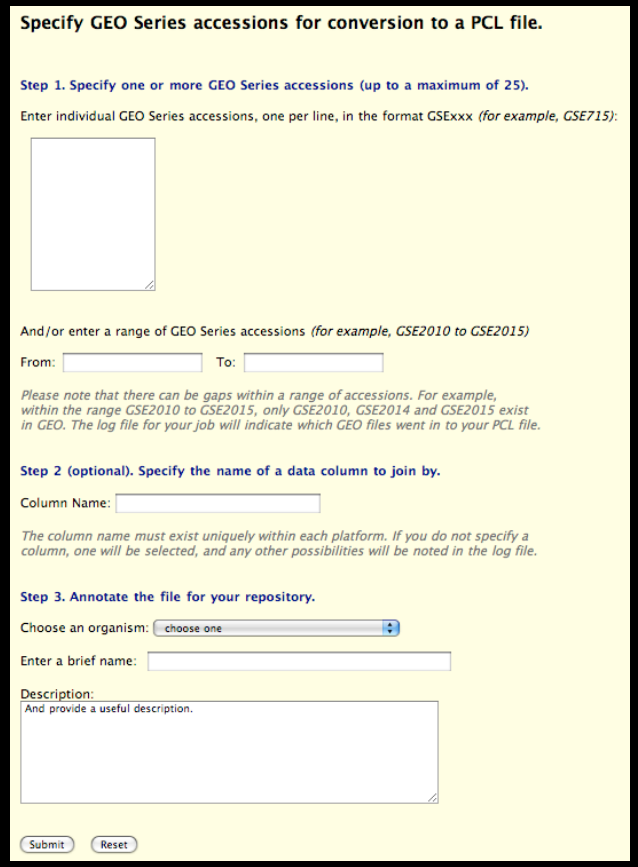
- Can be used to combine 2 pcl files from different sources into a single pcl file.
- Can be used to averaged repilcate assays (columns)
- Reporters that belong to the same gene can be combined into single row, based on a translation file (i.e. “synthetic genes”, discussed earlier).

Merge PCL Files



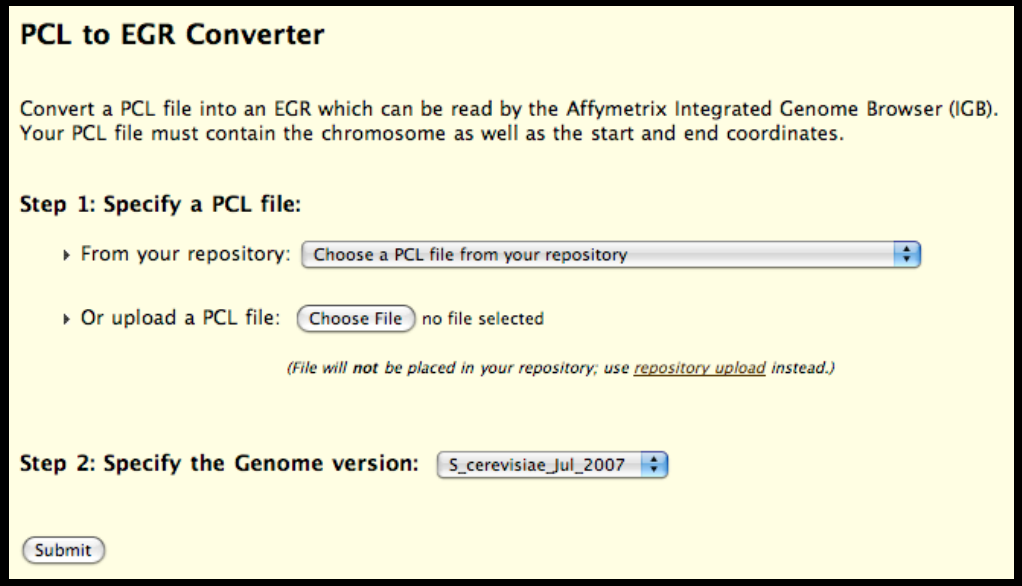
- Translation file:
  - Tab-delimited text file
  - First column: desired final identifier
  - Second column: desired final annotation
  - Third and subsequent columns: identifiers in the pcl files that should be collapsed to the identifier in the first column.
  - Data for identifiers not included in the translation file will not be collapsed
- Column averaging : replicate assays in the pcl files can be (mean/median)

GEO accession to PCL

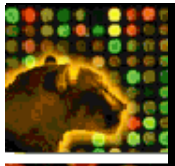


- Generate a PCL (pre-clustering) file from one or more NCBI GEO Series accessions
- Resulting PCL gets deposited to repository

PCL to EGR converter



- Currently only for PCL files where the chromosome, start, and end coordinates were retrieved
- Expression GRaph (EGR) files contain data for scored intervals.
- EGR files can be visualized using the Affymetrix Integrated Genome Browser (IGB)



Some References

Alter *et al*. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA (2000) vol. 97 (18) pp. 10101-6

Alter *et al*. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci USA (2003) vol. 100 (6) pp. 3351-6

Troyanskaya *et al*. Missing value estimation methods for DNA microarrays. Bioinformatics (2001) vol. 17 (6) pp. 520-5

Acknowledgements

- SMD
  - http://smd.stanford.edu
  - Catherine Ball, Gavin Sherlock, and staff
- PUMAdb, Lewis-Sigler Institute
  - http://puma.princeton.edu
  - Fan Kang, Laurie Kramer, Mark Schroeder, John Wiggins, Linda McMahan
- Affymetrix