PhenoGen Informatics

Published Monday, May 21, 2012, Version 2.6

Table of Contents

Overview

The PhenoGen Informatics website (http://phenogen.ucdenver.edu) is a comprehensive toolbox for storing, analyzing, and integrating microarray data and related genotype and phenotype data. The site is particularly suited for combining QTL and microarray data to search for "candidate" genes contributing to complex traits. In addition, the site allows, if desired by the investigators, storage and sharing of data. Investigators can conduct "*in-silico*" microarray experiments using their own and/or "shared" data.

The PhenoGen toolbox was originally created to facilitate interactions within the INIA consortium of investigators. In brief, the goals and purpose of the **INIA** (**I**ntegrative **N**euroscience **I**nitiative on **A**lcoholism, http://www.scripps.edu/cnad/inia/index.html) consortium is to identify the molecular, cellular, and behavioral neuroadaptions that occur in the brain reward circuits associated with the extended amygdala and its connections as a result of exposure to ethanol. Although PhenoGen web tools were initially created for the consortium members, the integrated tools described here can be used by the global scientific community.

Disclaimer

PhenoGen Informatics hopes that the tools made available will be useful to investigators in advancing the knowledge about genes through microarray research. However, since all of these tools rely on the information uploaded by various investigators and from various public databases (such as MGI, Ensembl, and NCBI), PhenoGen cannot guarantee the reliability of the data. Similarly, if any of these databases are not functioning properly, such malfunction is expected to affect the results of queries carried out on the PhenoGen website. In the past few years, availability of computational tools based on “Natural Language Processing” has considerably decreased the time needed for high-throughput literature searches. However, users should check the results of the "Literature Search” due to various caveats associated with extracting information out of biomedical literature using such computational tools (Hunter and Cohen, 2006, Molecular Cell, 21:589). The PhenoGen website tools use gene symbols and synonyms along with the user-defined keywords to search the PubMed database. For example, gene symbol “Cap1” (which could be either an official gene symbol for “adenylyl cyclase-associated protein 1” – MGI ID: 88262, or a synonym for “protease, serine, 8” – MGI ID: 1923810, or a short-form for “contraception-associated protein 1” – PubMed ID: 11105923) may pull abstracts related to any of these proteins and a gene symbol “Wars” (for tryptophanyl-tRNA synthetase) may get abstracts related to wars (fighting) rather than the actual gene. Please use the tools provided on the PhenoGen website with care and proper reflection and review of the output.

Citation for the PhenoGen Website

PhenoGen Website [Internet]. Aurora (CO): University of Colorado Denver, School of Medicine. PhenoGen Informatics, 2007 - [cited (insert date of access)]. Available from: http://phenogen.ucdenver.edu. Primary publication describing PhenoGen and use of tools available to investigate "candidate genes" for a complex trait: Bhave et al., 2007, BMC Genetics, 8:59.

Getting Started

Before you can use the PhenoGen website, you must **register** and set up a **user profile**. Your user profile provides details such as your name, email address, and the Principal Investigator you are working with. Your user profile can be modified at any time when you are logged into the website. After your registration is complete, you can use the website to create and analyze datasets and research gene lists. The website is available at **http://phenogen.ucdenver.edu/**

See "Registering an Account" and "Logging In and Out" on page 7 for details.

Minimum System Requirements

To successfully run the PhenoGen website on your browser, your computer must have:

- 1GB RAM
- 3.0 GHz CPU
- One of:
  - Firefox 2.0 or later
  - Internet Explorer 7.0 or later.
  - Safari 2.0 or later

Note: The features and functionality of the PhenoGen website may work with other browsers, but compatibility is not guaranteed, and the support provided for those browsers may be limited.

User Types

There are two types of users in the PhenoGen website:

- Basic User (default)
- Principal Investigator (PI)

See "Registering an Account" for details.

Basic User

Most users are basic users. These users can:

- Upload experiment arrays.
- View arrays and datasets.
- Create datasets.
- Download datasets.
- Access all of the available data analysis tools.
- Use all the tools available for researching genes.

Principal Investigator

The Principal Investigator (PI) is often the head of a lab and is responsible for granting permission to other users to view the arrays uploaded by researchers in the PI's lab. A user who is a Principal Investigator sees a Principal Investigator box on the Home tab after logging in. This box provides administrative functions for the PI. In addition to all of the functions available to basic users, the PI can:

- Approve array requests.
- Grant array access to an individual.
- Grant open access to array data.

PhenoGen Website Overview

The PhenoGen website shares experimental data with a worldwide community of investigators and provides a flexible, integrated, multi-resolution repository of neuroscience transcriptomic genetic data for collaborative research on genomic disorders.

The website provides a comprehensive system to organize, query, analyze, and retrieve high-throughput gene expression data, as well as providing users with computational tools for integrated analysis of neuroscience data, biomedical literature, gene functional annotations, and Quantitative Trait Loci (QTLs).

The PhenoGen website allows data to be classified as "Semi-public" or "Open Access". All of the information about the data uploaded at the PhenoGen website is visible to every registered user (see "Registering an Account" on page 6 for details). Registered users have full access to data that is classified as "Open Access" and do not need to obtain permission from the curator (Principal Investigator) of the data. "Semi-public" data can only be accessed and downloaded after the curator of the data grants a user permission to do so. Registered users can use the data for "in-silico" analysis or can download the data for analysis with their own statistical software.

The website also has nine pre-compiled “Public” microarray datasets that can be used and downloaded by all registered users for gene expression analysis, including correlating with user-provided phenotype data. These datasets include inbred and recombinant inbred mice and rat strains.

The PhenoGen website allows you to:

- Upload microarray raw data into a MIAME-compliant database.
- Upload gene lists.
- Share data with other investigators around the world.
- Search literature and save results.
- Translate gene identifiers to and from multiple databases.
- Determine phenotypic QTLs using BXD recombinant inbred mice, HXB/BXH recombinant inbred rats, or LXS recombinant inbred mice.
- Match physical location of genes of interest and their eQTL to phenotypic QTLs.
- Correlate gene expression with a phenotype.

You can also perform:

- Microarray data quality control analysis and normalization.
- Data filtering (noise filtering).
- Statistical analyses, including the most common statistical tests and permutations.
- Promoter analysis (transcription factors).
- Queries about genetic variations (e.g. SNPs or polymorphisms) in the transcripts of interest.

Website Process Flow for Microarray Analyses

Analyze Microarray Data

The process flow for a microarray analysis is:

Dataset Creation

Upload microarray data. If you have microarray data from a lab experiment, you can upload it into a MIAME-compliant database that is part of the PhenoGen website.

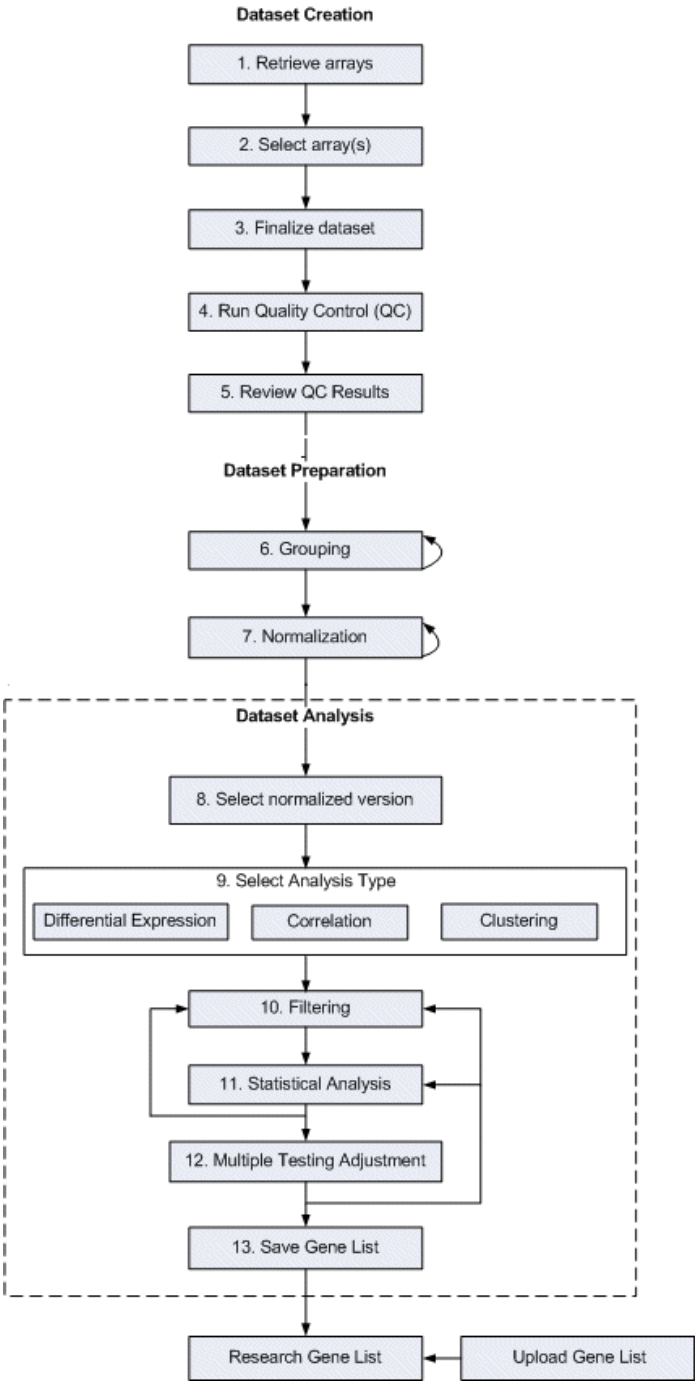
1. Retrieve arrays.
2. Select and merge arrays from the data repository.
3. Finalize dataset.
4. Run quality control measures on the merged arrays.
5. Review quality control results.

Dataset Preparation

6. Group the arrays based on your hypothesis (e.g., disease vs. control).
7. Normalize the dataset.

Dataset Analysis

8. Select a single normalized version.
9. Select analysis type.
10. Filter genes.
11. Perform statistical analysis.
12. Perform a multiple testing adjustment
13. Save gene lists.



Research Gene Lists

Do one of the following to enter a gene list in PhenoGen:

- Generate a gene list from microarray analysis (see the preceding Analyze Microarray Data section.)

  OR

- Upload a gene list if you have an existing gene list to interpret.

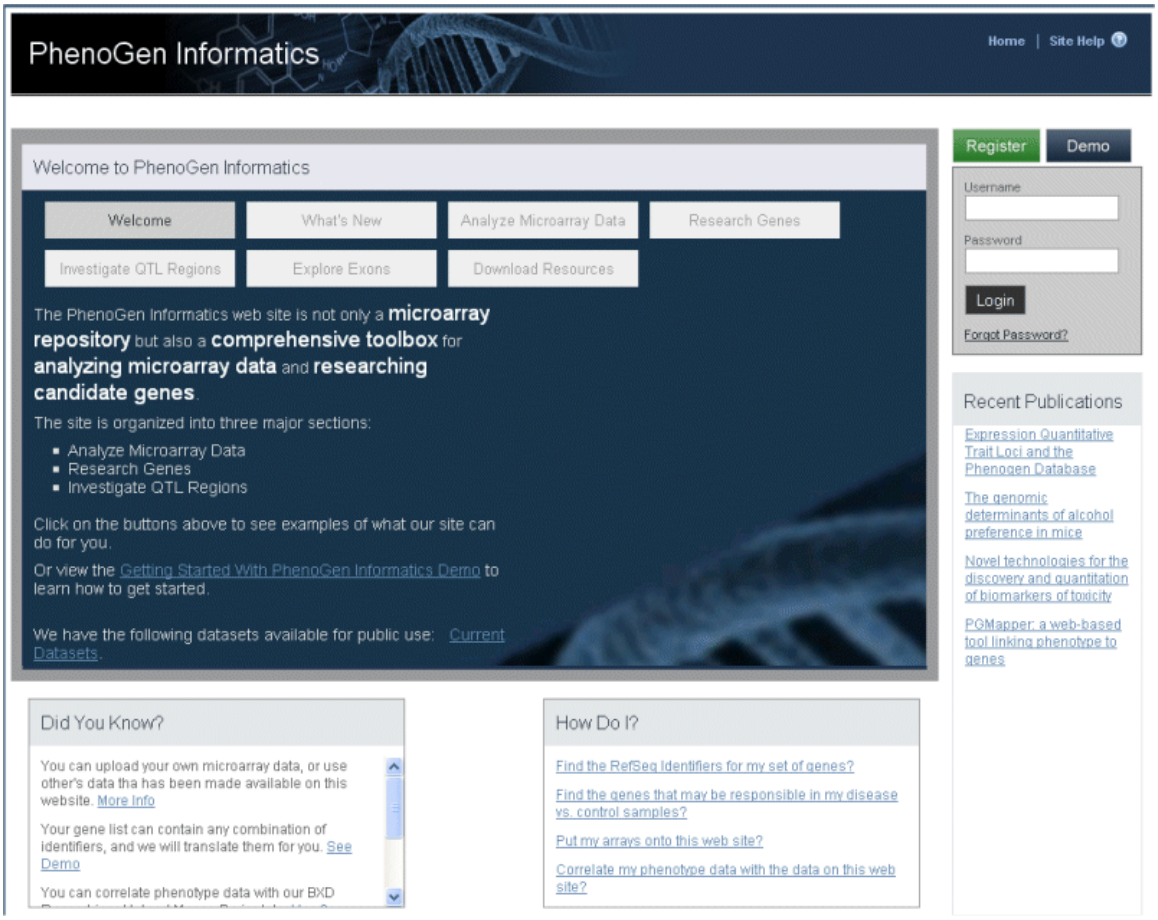
When your gene list is on the website, use the annotation, QTL, literature search, and promoter analysis tools to help interpret your list of candidate genes.

Website Home Page

The PhenoGen Home page allows you to :

- Review the types of tools available to registered users.
- Register to use the website.
- Log in to the website.
- Retrieve a forgotten password.

The two boxes at the bottom, *Did You Know* and *How do I,* provide quick links to pertinent topics and instructions for performing specific tasks. The *Recent Publications* box provides links to documents that relate to the PhenoGen database, gene analysis, and other topics of interest.

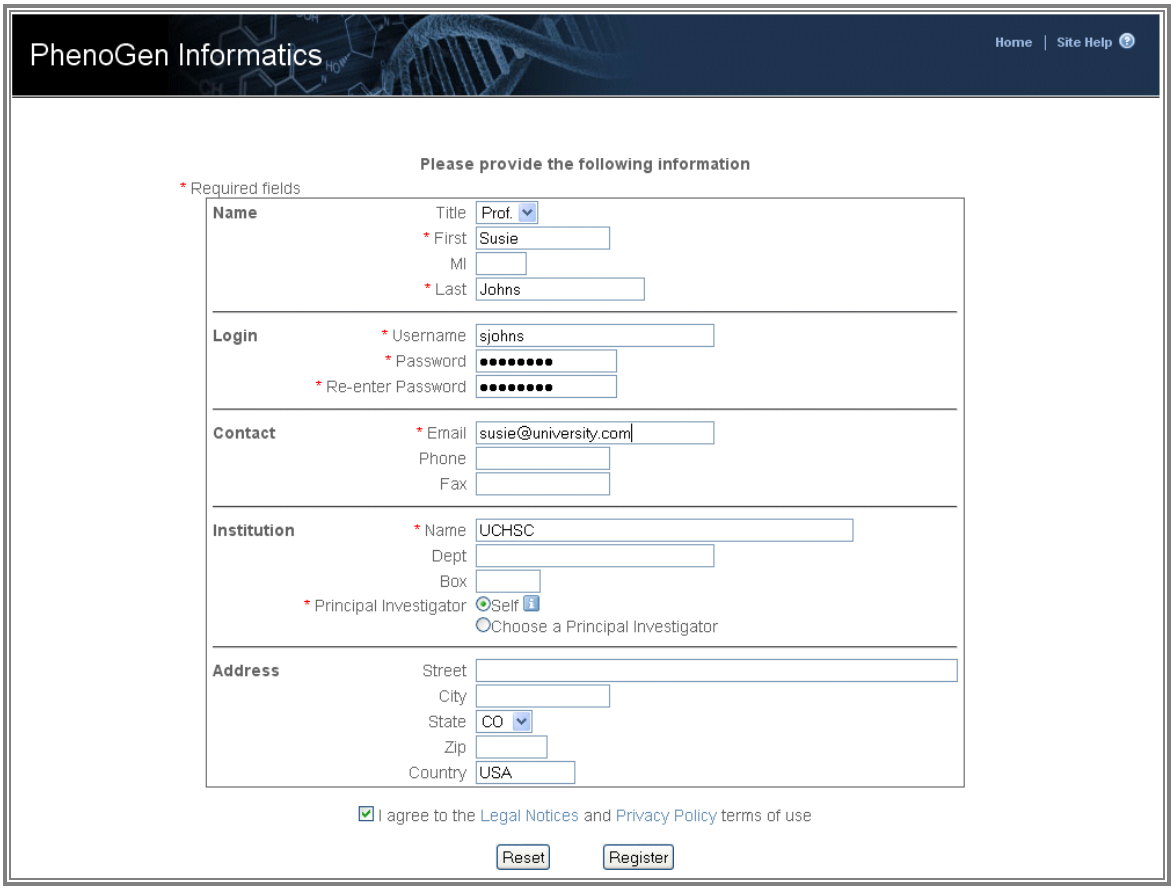


Registering an Account

The Registration page must be completed before you can log into the PhenoGen website.

1. Click **Register** on the *Home* page. The *Registration* page displays.

   Note: Required fields have an asterisk *.



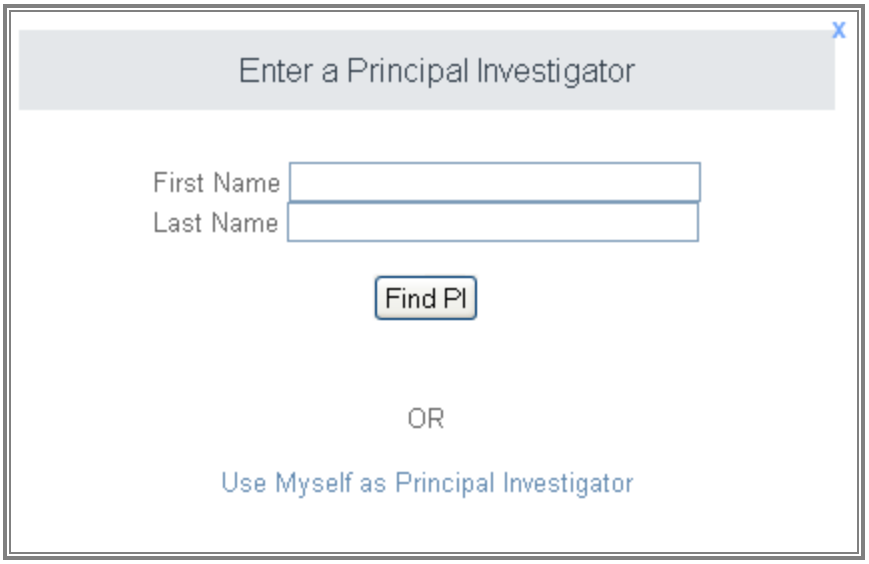
2. Select your **Title** and enter your **First Name**, **Middle Initial**, and **Last Name**.
3. Enter a **Username**. If the username you enter is the same as an existing name, an error message displays when you try to register.
4. Enter a **Password**, then re-enter the password. Your password must be 6 to 16 characters and must contain numbers, letters, and special characters (~!@#$%^&*()+).

   Note: Your Username and Password allow you to log into the website.

5. Enter your contact **Email**, **Phone**, and **Fax**.
6. Enter the **Name** of your institution, your **Department**, and your **Box** number.
7. Select **Self** if you are the principal investigator, or select the **Choose a Principal Investigator** option.

   Choose a PI

   1. Enter the **First** and **Last Name** of the PI, and click **Find PI**.



2. Click **Choose this PI** if the system finds the Principal Investigator you want.

8. Enter your **Address**.
9. Select **I agree to the Legal Notices and Privacy Policy terms of use**.
10. Click **Register**. A pop-up displays with the terms of the PhenoGen website. If you agree to the terms, click **OK** to send your request for registration. Click **Cancel** if you do not agree to the terms.

If your registration request is successful, a page displays that informs you that your submission was successful and the submission will be reviewed within 24 hours.

See "Updating Your Profile" for instructions on changing your registration information. The only detail you cannot change is your username.

Logging In and Out

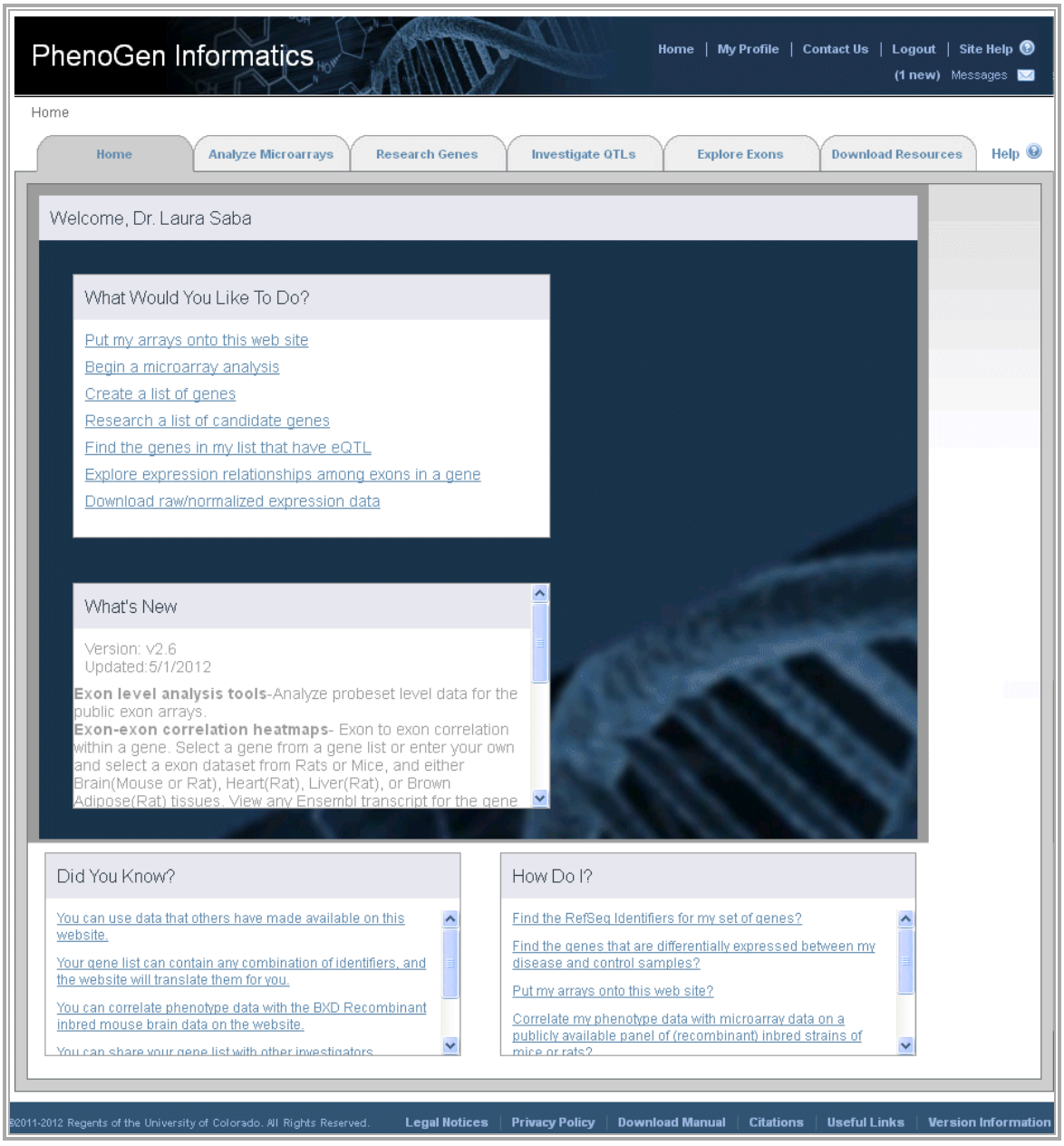
Logging Into the Website

Before you can log into the PhenoGen website, you must register and receive approval for your registration. After your registration is approved:

1. Open your internet browser.
2. Type **http://phenogen.ucdenver.edu** in the location bar, and press **Enter** on your keyboard. The PhenoGen website displays.
3. Enter your **Username**.
4. Enter your **Password**.
5. Click the **Login** button below the Password field to log in. The *Home* tab displays.

Your Home Page

After you log in, the *Home* tab is personalized. A *What would you like to do?* box displays task options you might want.



Logging Out of the Website

Log out of the website when you are done entering data, creating datasets, and analyzing data. When you log out, the website knows that you are finished and closes your connection to the website. You can log in again at any time.

- Click **Logout** at the top right corner of the page.

Updating Your Profile

You can update the information you provided in the Registration page.

1. Log into the website.
2. Click **My Profile** at the top right. The *Registration* page displays.

   Note: Required fields have an asterisk *.
3. Update your information as required. You cannot change your username.
4. Click **Update** to update your information, or **Reset** to return all the fields to their original values.

Using the PhenoGen Website

The PhenoGen website has a number of conventions that are common throughout all the pages and tabs.

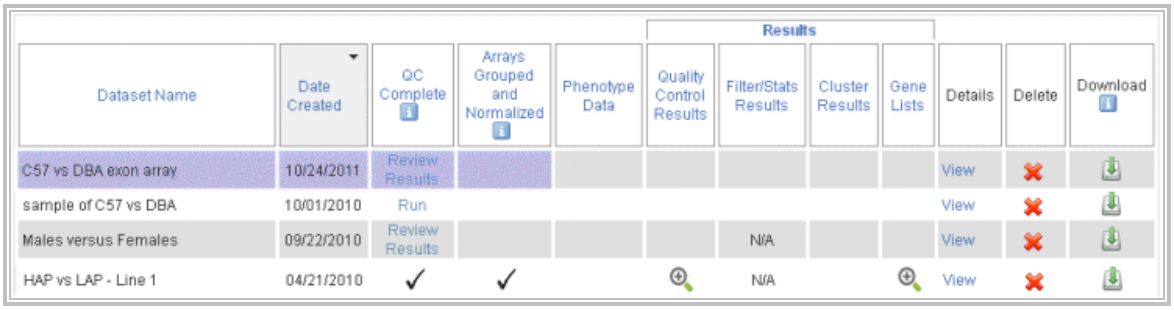
Icons

Icons in the tables on some pages show the actions that you can take for specific data:

Icon	Action
Delete	Click the **Delete** icon to delete the row in which the icon displays. The Delete icon allows you to delete datasets, gene lists, and QTL Regions. It also displays on many detail pages, such as the Grouped and Normalized page or the tabs for a specific gene list, to allow you to delete specific items within a dataset or gene list.
Download	Click the **Download** icon to download the data for the row in which the icon displays. The Download icon allows you to download datasets, gene lists, and QTL Regions.
View Details	Click the **Magnify** icon to display details for the item on the current page, such as a specific dataset or gene list.
Help	Click the **Help** icon at the top of the page beside the words "Site Help" or on a specific tab or page beside the word "Help". When you click Site Help, the full version of the help displays. When you click Help, help for the specific page or tab displays.
Information	Hover over the **Information** icon for an explanation of the item the icon is beside.
plus minus	Click the **plus** or **minus** icons to show and hide more information.
sort A-Z sort Z-A	Click the arrows to change the sort order in a column.

Rows

Rows in tables change to purple when you mouse over them. Click on a row to select it. Click the **View** link to display row details.



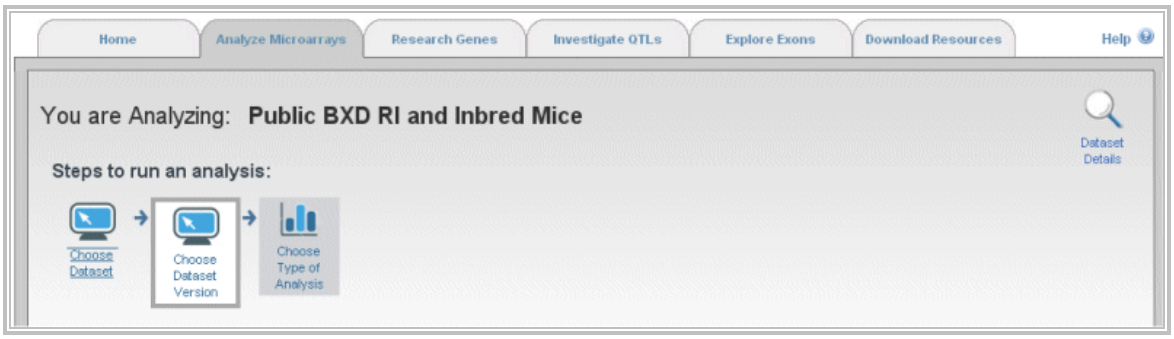
Depending on your browser settings, items you click may display in the current browser window, a new browser window, or a new tab.

Links

Many pages have links below the "Steps..." and at the top right (Note the *Dataset Details* link at the top right in the following screen shot). The links may allow you to:

- Quickly select new data without clicking a tab and starting again at the beginning of the selection process.
- View details of your current data.
- Download the current data.
- Perform other actions, like creating a new normalized version or creating a new dataset.

Link options are customized for each page. For example, in the analyzing datasets process, a **Choose Dataset** link displays below the *Steps to run an analysis* and allows you to change datasets.



Breadcrumbs

Above the tabs on most pages, "breadcrumbs" display. Breadcrumbs are a navigation technique that show you the page you are on and allow you to click back to previous pages. Breadcrumbs look similar to:

Home > *Analyze Microarrays* > *Create Group*

Where *Create Group* is the current page, and *Home* and *Analyze Microarrays* are both links back to those pages.

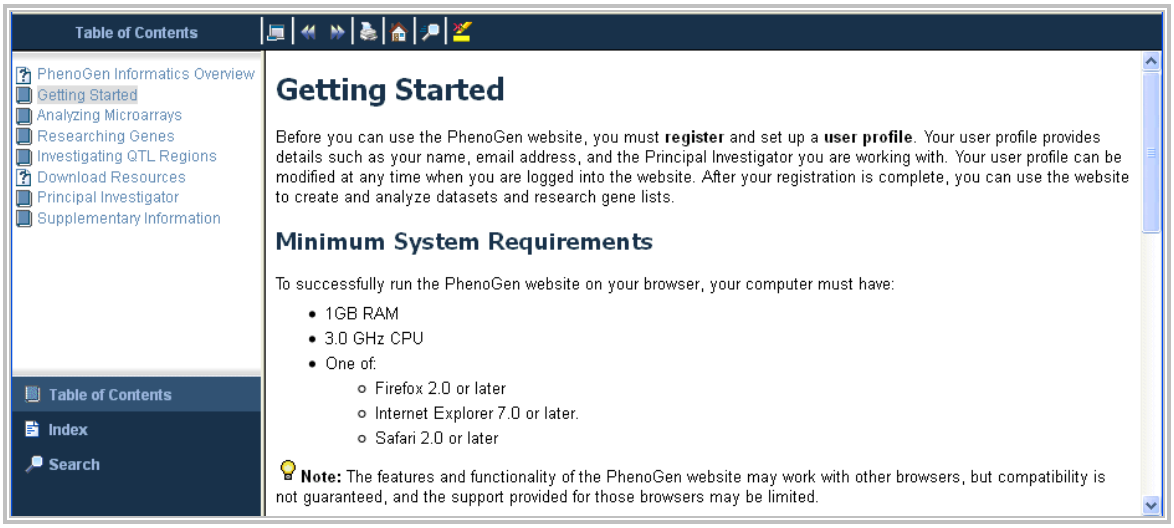
Using the Online Help

The PhenoGen website has online help to assist you. You can open the full online help or task-based help for each page in the PhenoGen website. Most instructional pages in the help contain links to related topics in a **See Also** section. You can also download the latest PDF version of the PhenoGen User Guide.

Full Version of Online Help

- Click the **Site Help** link that displays at the top right of each page.

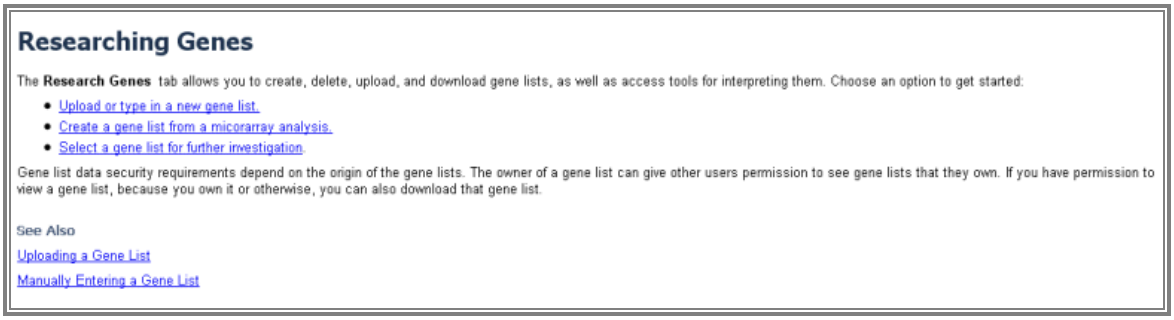
Use the table of contents, the index, or the search to find the topics that you want.



Task-based Version of Online Help

Task-based help is pertinent to a specific page in the PhenoGen website.

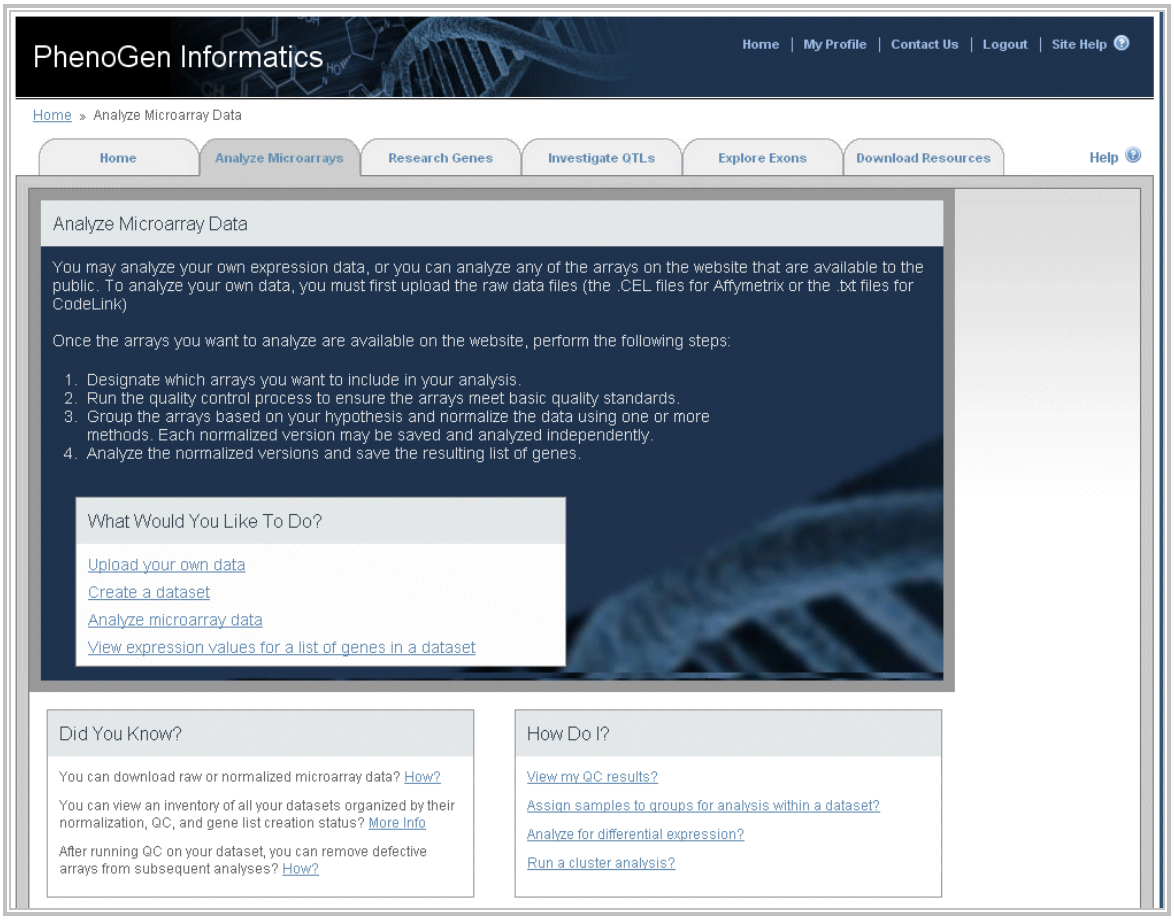
- Click the **help** link on any page to open the task-based topic related to the current page. The following image is an example of a task-based topic.



Analyzing Microarrays

The **Analyze Microarrays** tab allows you to upload your own arrays (the raw data files: .CEL files for Affymetrix or .txt files for CodeLink) or analyze existing microarray data. After the microarray data you want is available, you can:

1. Designate which arrays you want to include in your analysis.
2. Run the quality control process to ensure the arrays meet basic quality standards.
3. Group the arrays based on your hypothesis, and normalize the data using one or more methods. Each normalized version can be saved and analyzed independently.
4. Analyze the normalized versions and save the resulting list of genes.



Choose an option to get started:

- Upload my own data. See "Uploading Your Arrays".
- Create a dataset. See "Creating Datasets".
- Analyze microarray data. See "Viewing Datasets".
- View expression values for a list of genes in a dataset. See "Viewing Gene Expression Data".

Viewing Datasets

You can see the collection of all your datasets at any time. The page that displays your datasets shows you the stage that each dataset is in (e.g., quality control completed, grouped and normalized, etc.)

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.

At the top of the page, you can click the **Create Dataset** option if you want to retrieve and select arrays and finalize them into a new dataset, or click the **Upload Arrays** option to upload your own arrays and create a dataset. The page provides four grouped and normalized "Public" datasets for you to analyze and save new gene lists. See "Public Datasets" for information about public datasets.

Another table displays after the Public Datasets table and shows your "Private" datasets. After you finalize a dataset, it becomes part of the *My Private Datasets* table, where your progress on that dataset is denoted in the columns:

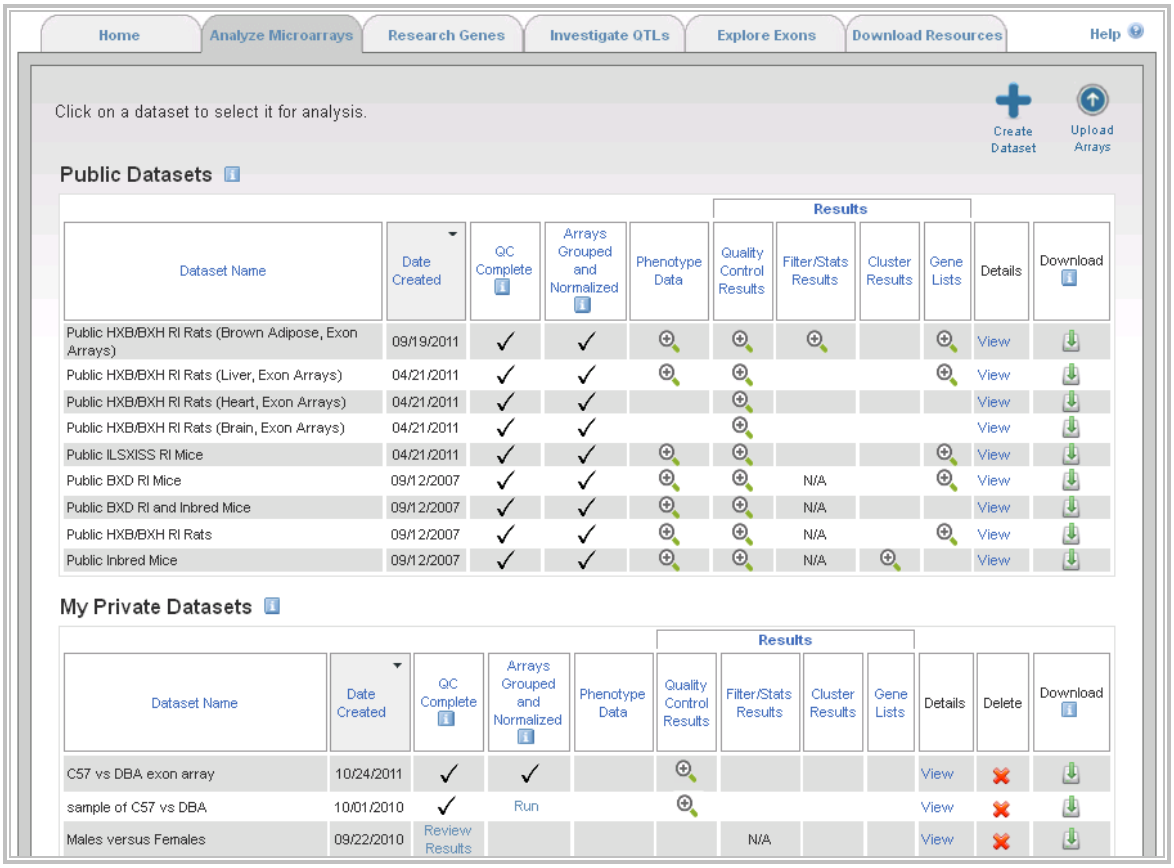
- **QC Complete**: The word "Run" displays in the QC Complete column until you perform quality control checks. At that time, "Review Results" displays in the QC Complete column. You must review your quality control results and approve them before a checkmark displays in this column.
- **Arrays Grouped and Normalized**: The word "Run" displays in the Arrays Grouped and Normalized column until you create a group and normalize the data based on that grouping. You can create multiple array groupings and normalize each group many different ways. Each normalized grouping is saved as a new dataset "version". After you create and normalize your grouped arrays, a checkmark displays in this column.
- **Phenotype Data**: A magnifying glass displays in this column if there is phenotype data for the dataset.

Results section

- **Quality Control Results**: A magnifying glass displays in this column if there are quality control results for the dataset.
- **Cluster Results**: A magnifying glass displays in this column if there are cluster results for the dataset.
- **Gene Lists Saved**: A checkmark displays in this column after you analyze the normalized dataset and save the resultant gene list.

You can:

- Click a dataset to view that dataset in its current state of processing.
- Click the **View** link in the **Details** column to view dataset details such as name, description, organism, arrays in dataset, and more. See "Viewing Dataset Details" for more information.
- Click the **Delete** icon  to delete a dataset.
- Click the **Download** icon  to download a dataset.



Public Datasets

The Public datasets available for analysis on the PhenoGen website are pre-compiled groupings of gene expression data for various strains of inbred and recombinant inbred mice and rats. These datasets are available for all types of analysis by any registered users but may be most useful for performing correlation analysis with phenotype data. These datasets are normalized using the most common normalization techniques and have already had quality control checks run. Additionally, datasets created using the Affymetrix Exon arrays have been adjusted for batch effects using an empirical Bayes method (Johnson et al 2007). The normalized data or raw data can be downloaded from the Download Resources page.

Inbred Mice

The whole brain gene expression dataset for the inbred mice includes 20 inbred strains. Each strain has four to seven biological replicates for a total of 90 individual arrays. The whole brain mRNA for each naive 10-12 week old male mouse was hybridized to a separate array, i.e., no pooling of samples.

The inbred mouse data was normalized nine different ways. Five of the normalization methods are available on the website. For the other four versions, a probe mask was created to eliminate probes whose sequences did not match to the NCBI m37 Build, matched the genome in multiple places, or harbored a SNP between any of the 19 strains where genotype data is available at the Imputed Genotype Resource from the Jackson Laboratory; http://cgd.jax.org/datasets/popgen/imputed.shtml (129P3/J is not available). Entire probe sets were eliminated if less than four associated probes remained. The version using the probe mask and the RMA normalization method is the RECOMMENDED version.

BXD Recombinant Inbred Mice

The whole brain gene expression dataset for the BXD recombinant inbred mice includes 30 recombinant inbred strains and the two parental strains (C57BL/6J and DBA/2J). Each strain has four to seven biological replicates for a total of 172 individual arrays. The whole brain mRNA for each naive 10-12 week old male mouse was hybridized to a separate array, i.e., no pooling of samples.

The BXD data was normalized nine different ways. Five of the normalization methods use all of the probes in the dataset. A probe mask was created for the other four versions to eliminate probes whose sequences did not match to the NCBI m37 Build, matched the genome in multiple places, or harbored a SNP between any of the 19 inbred mouse strains included in the public dataset (according to Imputed Genotype Resource from the Jackson Laboratory; http://cgd.jax.org/datasets/popgen/imputed.shtml). Entire probe sets were eliminated if less than four associated probes were eliminated. The version using the probe mask and the RMA normalization method is the RECOMMENDED version.

For the eQTL analysis of this data set, a slightly different mask was used. Instead of eliminated probes with SNPs between the 19 inbred strains, probes were eliminated if they contained a known SNP between the two BXD parental strains, based on whole genome sequence data from the Sanger Institute (Keane et al 2011). Expression values were normalized and summarized into probesets using RMA. MAS5 was used to evaluate if expression level measurements were above background noise (present, absent, or marginal). If a probeset did not have at least one present call throughout all samples, the probeset was dropped from the data set. Of the 41,581 probesets retained after masking, 30,031 probesets remained after filtering by present/absent calls. Data were thoroughly examined for batch effects related to processing. The microarrays were run over a year and a half period, resulting in 15 batches. Both batches and strains contribute to non-random data distribution and a new method for removing batch effects, while retaining strain effects, was used (personal communication, Evan Johnson, Boston University) on the set of 30,031 probesets detected above background. This method combines a simple rank test and a Bayesian hierarchical framework similar to the empirical Bayes method, *Combating Batch Effects When Combining Batches of Gene Expression Microarray Data* (ComBat) (Johnson et al., 2007). This version of the data is available in the Download Resources section.

BXD Recombinant Inbred and Inbred Mice

This expression data set represents a combination of the two datasets previously mentioned. In this dataset there are a total of 50 strains (C57BL/6J and DBA/2J are in both of the previous sets) and 253 individual arrays. See the preceding topic for details on "masked" versions.

LXS Recombinant Inbred Mice

The whole brain gene expression dataset for the LXS recombinant inbred male mice includes 59 recombinant inbred strains (one strain (LXS49) was eliminated due to unresolved questions about true strain origin) and two parental strains (ILS and ISS). Each strain has three to six biological replicates, for a total of 342 individual arrays that passed quality control standards. In addition, to control for batch effects, C57BL/6J mice were hybridized to arrays and included in every batch (35 individual arrays), and DBA/2J mice were included in a few of the final batches (9 arrays). The whole brain mRNA for each naive 10-12 week old male mouse was hybridized to a separate Affymetrix Mouse Exon Array 1.0 ST, i.e., no pooling of samples.

Individual probes were eliminated prior to normalization if their sequence did not match any part of the NCBI m37 Build of the mouse genome, if their sequence matched multiple locations in the mouse genome, or if the location in the genome that the probe did match contain a SNP between any of the 19 strains in the public Inbred Mice dataset where genotype data is available at the Imputed Genotype Resource from the Jackson Laboratory; http://cgd.jax.org/datasets/popgen/imputed.shtml (same mask that is implemented on PhenoGen). Entire probe sets were eliminated if less than three of the original probes remained after filtering. Arrays were examined for quality, and arrays that did not meet quality standards were eliminated.

Data from individual probes was normalized using RMA and summarized either into the full set of transcript cluster or the core set of transcript clusters. In addition, RMA values for the full set of individual probe sets is available for download from the resource page, but is not available for analysis on PhenoGen at this time.

Each data set was adjusted for batch effects using the empirical Bayes method outlined by Johnson et al (2007). After batch effects adjustment, C57BL/6J and DBA/2J arrays were dropped from the data set. The version using the probe mask and the RMA normalization method on the core transcript clusters is the ***recommended*** version and was used for calculation of eQTLs.

HXB/BXH Recombinant Inbred Rats

The whole brain gene expression dataset for the HxB/BxH recombinant inbred rats on the CodeLink Whole Genome Rat Array includes data from 26 recombinant inbred strains, the two parental strains (SHR/Ola and BN-Lx/Cub), and the SHR-Lx/Cub strain. The whole brain mRNA of four to seven naive 12-14 week old male rats from each strain were hybridized to separate CodeLink Whole Genome rat arrays (one rat per array) for a total of 139 arrays.

In addition to the five normalization versions available on the website, an "eQTL version" of the dataset that was used for all HXB/BXH rat eQTL calculations is available. This version was obtained by first removing probes from the datasets if they were one of the negative or positive controls placed on the array by the manufacturer. Next, individual values were eliminated based on the quality flags assigned by the CodeLink Expression Analysis Software. Values were eliminated if they were flagged as M (spot was identified to be defective through image inspection at manufacturing), C (spot has a high level of background contamination), I (spot has an irregular shape), or S (spot has a high number of saturated pixels). Values were retained if they were flagged G (spot is good) or L (spot is below local background noise). Also, to be able to take the log base 2 transformation of the background-adjusted intensity values, all background-adjusted intensity values below zero were replaced with the value 0.00001. The data was then normalized using a cyclic LOESS procedure executed in R to account for the missing intensity values.

The HXB/BXH recombinant inbred panel also has four data sets available on transcription levels from the Affymetrix Rat Exon array. Data was collected on whole brain, left ventricle (heart), liver, and brown adipose tissue (BAT) of 21 HXB/BXH RI strains (only 19 RI strains included in the BAT tissue data set) and 6 related inbred strains. Each strain has three to four biological replicates for a total of 108 individual arrays from brain, 105 arrays from heart, 106 arrays from liver, and 96 arrays from brown adipose tissue that passed quality control standards. The mRNA for each naive 10 week old male rat was hybridized to a separate Affymetrix Rat Exon Array 1.0 ST, i.e., no pooling of samples.

Individual probes were eliminated prior to normalization if their sequence did not match any part of the RGSC version 3.2 of the rat genome, if their sequence matched multiple locations in the mouse genome, or if the location in the genome that the probe did match contain a SNP between the Brown Norway (BN/SsNHsdMcwi) inbred strains (reference strain) and the spontaneously hypertensive rat (SHR/OlaIpcv) strain that was recently sequenced (Atanur et al 2010) using next generation sequencing or a SNP detected in DNA sequencing of the BN-Lx/CubPrin and SHR/OlaIpcvPrin strains (same mask that is implemented on PhenoGen). DNA sequence data for the BN/SsNHsdMcwi and SHR/OlaIpcv was downloaded directly from the Ensembl ftp site at: ftp://ftp.ebi.ac.uk/pub/databases/ensembl/snp/rat/shr/.

For the 4,022,111 original probes, 604,601 were removed (472,072 did not map uniquely to the genome; 132,529 probes contained a SNP). Entire probe sets were eliminated if less than three of the original probes remained after filtering. Arrays were examined for quality and arrays that did not meet quality standards were eliminated.

Data from individual probes was normalized using RMA and summarized either into the full set of transcript cluster or the core set of transcript clusters. In addition, RMA values for the full set of individual probe sets is available for download from the resource page, but is not available for analysis on PhenoGen at this time.

Each data set was adjusted for batch effects using the empirical Bayes method outlined by Johnson et al. (2007). The version using the probe mask and the RMA normalization method on the core transcript clusters is the ***recommended*** version and was used for calculation of eQTLs.

References

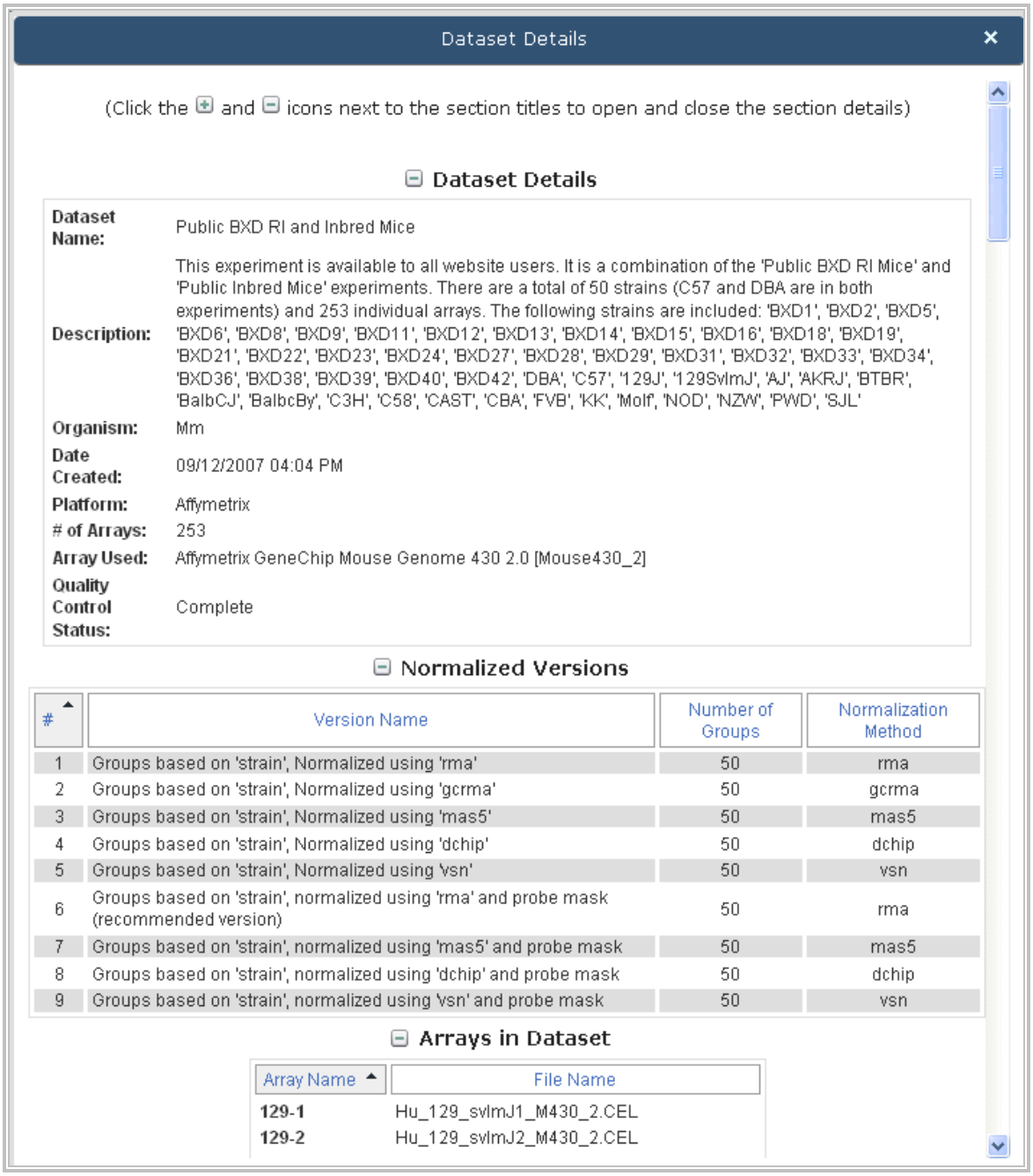
1. Johnson WE, Li C, and Rabinovic A (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1): 118-127.

2. Atanur SS, Birol I, Guryev V, Hirst M, Hummel O, Morrissey C, Behmoaras J, Fernandez-Suarez XM, Johnson MD, McLaren WM, Patone G, Petretto E, Plessy C, Rockland KS, Rockland C, Saar K, Zhao Y, Carninci P, Flicek P, Kurtz T, Cuppen E, Pravenec M, Hubner N, Jones SJ, Birney E, Aitman TJ (2010). The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Research 20(6):791-803.

Viewing Dataset Details

You can view details for datasets after you finalize them, before, during, and after quality control, grouping, normalization, and after you save the results as a gene list.

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.
3. Click the **View** link in the **Details** column beside a dataset to view the details. You can also view the dataset details when you click the magnifying glass beside the dataset name.



The *Dataset Details* page displays information about the dataset, such as dataset name, organism, date created, and quality control status. It shows the normalized versions and the arrays in the dataset.

Uploading Your Arrays

The arrays available for analysis on the website using the PhenoGen web-based submission tool.

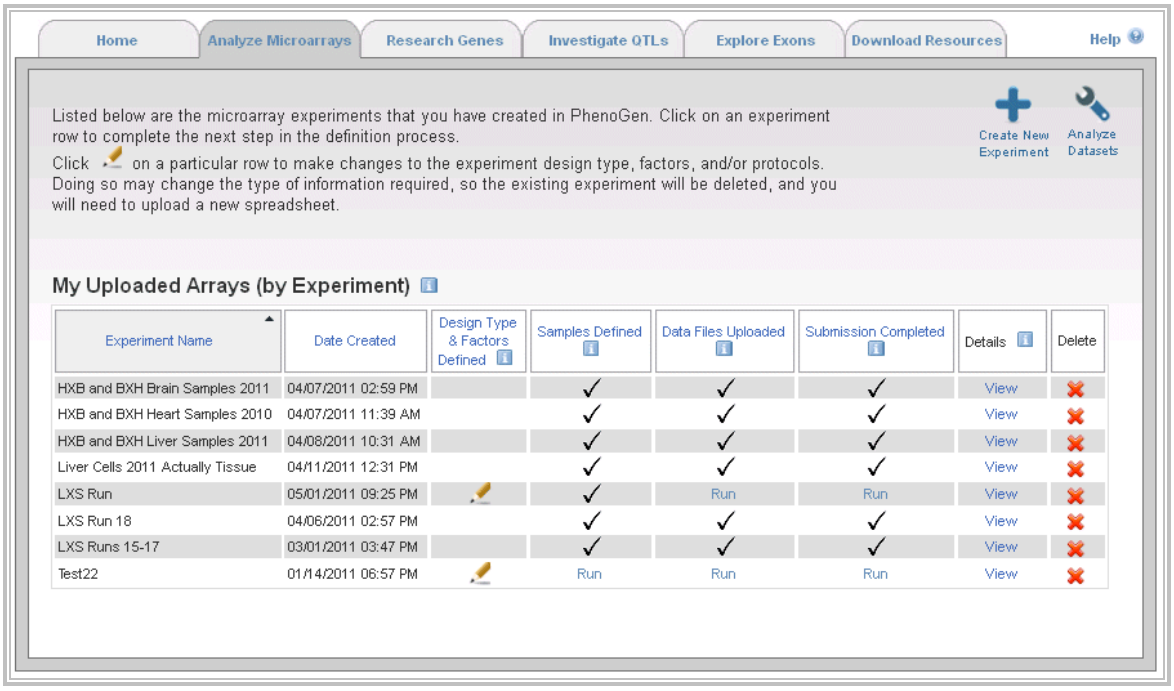
When you are logged into the PhenoGen website, you can upload microarray data (arrays). Each time an array is uploaded, it is automatically assigned to the Principal Investigator associated with the user who uploads the data.

Uploading an Array

1. Click the **Home** tab, then click **Put my arrays onto this web site** in the *What would you like to do* section. The *Upload Arrays* page displays and shows a list of Microarray Experiments you have entered.

   OR

2. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
3. Click **Upload your own data** in the *What would you like to do* section. The *Upload Arrays* page displays and shows a list of Microarray Experiments you have entered.
4. Click **Create New Experiment**.



Note: If you do not complete all the steps at once, follow steps 1-3, then click the experiment name, and the creation process begins where you left off. Click the **pencil** icon to edit the whole experiment. See "Editing Your Experiments" for details.

5. Enter an **Experiment Name**.
6. Enter a **description** of the experiment.

   Note: You should enter quality control information for any microarrays that failed quality control in the **Experiment Description** field.
7. Choose one or more **Design Type(s)** and **Experimental Factor(s)**. Click **View** beside each for a description.
8. Click **Next**.

Certain protocols are required to enter your data, and there are public protocols available to everyone for Extraction, Labeling, Hybridization, and Scanning. If you need protocols that are not listed, click **Create New** in the appropriate section to enter a new protocol name and description.

Note: You can delete your private protocols if they have not been used in an experiment.

9. Choose the appropriate protocol from each applicable section.
10. Click **Next**. A message displays that informs you to continue, you must download and enter data into, an Excel spreadsheet.
11. Close the message.
12. Click **Download Empty Spreadsheet**. An Excel spreadsheet with the same name as your experiment is downloaded.
13. Open the file if you plan to enter your data immediately, or save the file.

    Note: If you save, the file saves to the default download location for your current browser. See the browser help for information on downloads.

Downloading the Spreadsheet and Creating Samples

Enter a row for each sample in your experiment.

Important! You must have Internet access when filling out the spreadsheet.

14. Enter the **Hybridization Name** and **Sample Name**.
15. Fill in as much information as you can for each row. Data must be entered in the following fields to ensure arrays display correctly in the PhenoGen website. Optional fields and sections are not listed.

    Basic Sample Properties

    - **Organism**
    - **Sex**
    - **Organism Part** – E.g., "brain" for whole brain or "brain, left" for left hemisphere of brain.
    - **Sample Type** – In most cases this will be "frozen".
    - **Development Stage** – The developmental stage of the organism's life cycle during which the biomaterial was extracted.
    - **Genetic Modification** – If you do not find the appropriate modification, select **Other**, then enter details. Note that the drop-down list provides options that become searchable in the PhenoGen website *Browse Arrays* page.

    Protocol Details

    - **Extract Name**
    - **Extraction Protocol** – The procedure of extracting nucleic acid from the biomaterial.
    - **Labeled Extract Name**
    - **Labeled Extract Protocol** – The BioSample after labeling for detection of the nucleic acids.

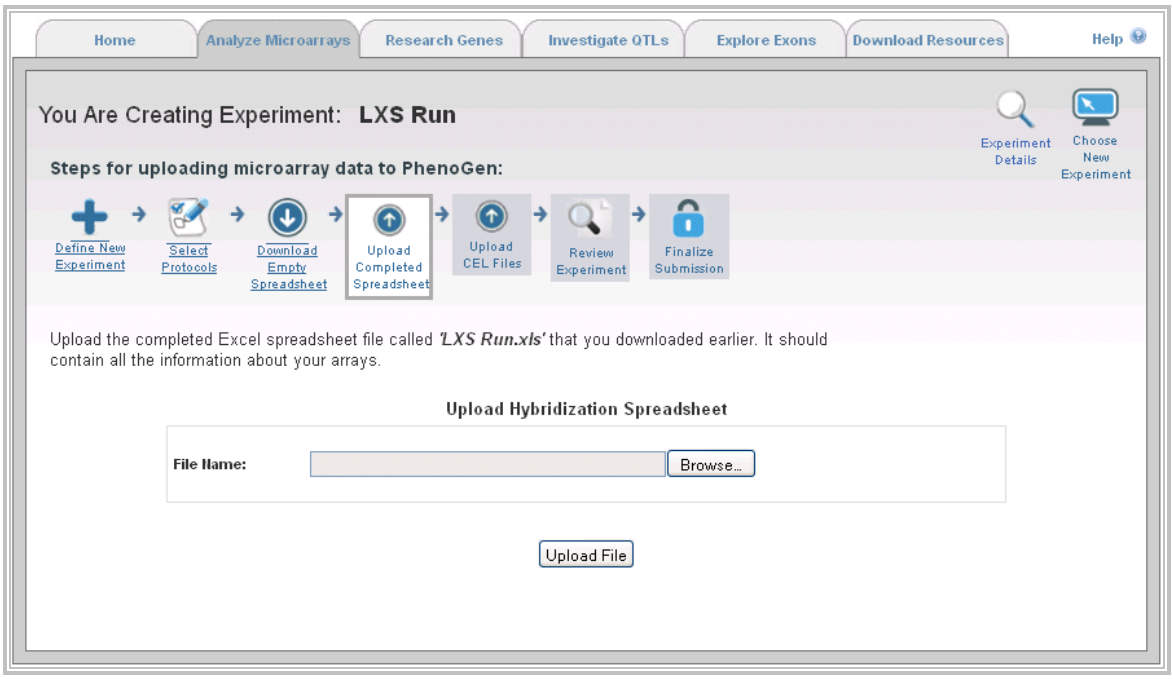
Hybridization Details

- **Array Design Name** – The array platform that was used.
- **Hybridization Protocol** – The process of incubating one or more labeled extracts with an array.
- **Scanning Protocol** – The process of applying a solvent (e.g. water) or a solution (e.g. SSC/SDS) to a BioMaterial or an array to remove impurities or unwanted compounds.

16. Save the file in Microsoft Excel 97 format with the same name as the experiment name (this is the default file name).

Uploading the Completed Spreadsheet

17. Open the PhenoGen website. If you are not already on the *Upload Arrays* page, click the **Analyze Microarrays** tab, then click **Upload your own data** in the *What would you like to do* section.
18. Click **Run** in the *Samples Defined* column for your experiment.
19. Click **Browse** to find your Excel file.

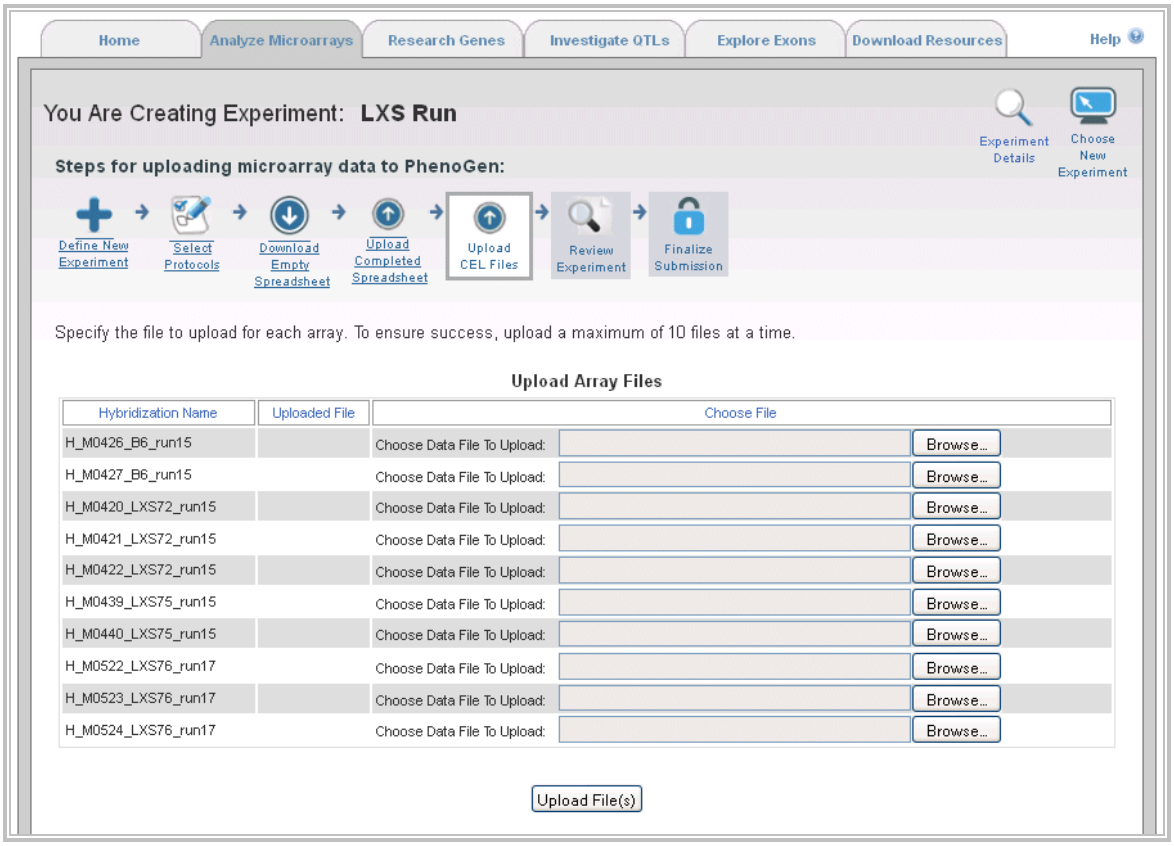


20. Click **Upload File**. A message displays that states if the spreadsheet upload was successful and also lists warnings and errors (if any). If you have errors, you must fix the errors and upload the spreadsheet again.

Uploading Data Files

After you upload the completed spreadsheet, you must upload array files that correspond to each hybridization in the spreadsheet.

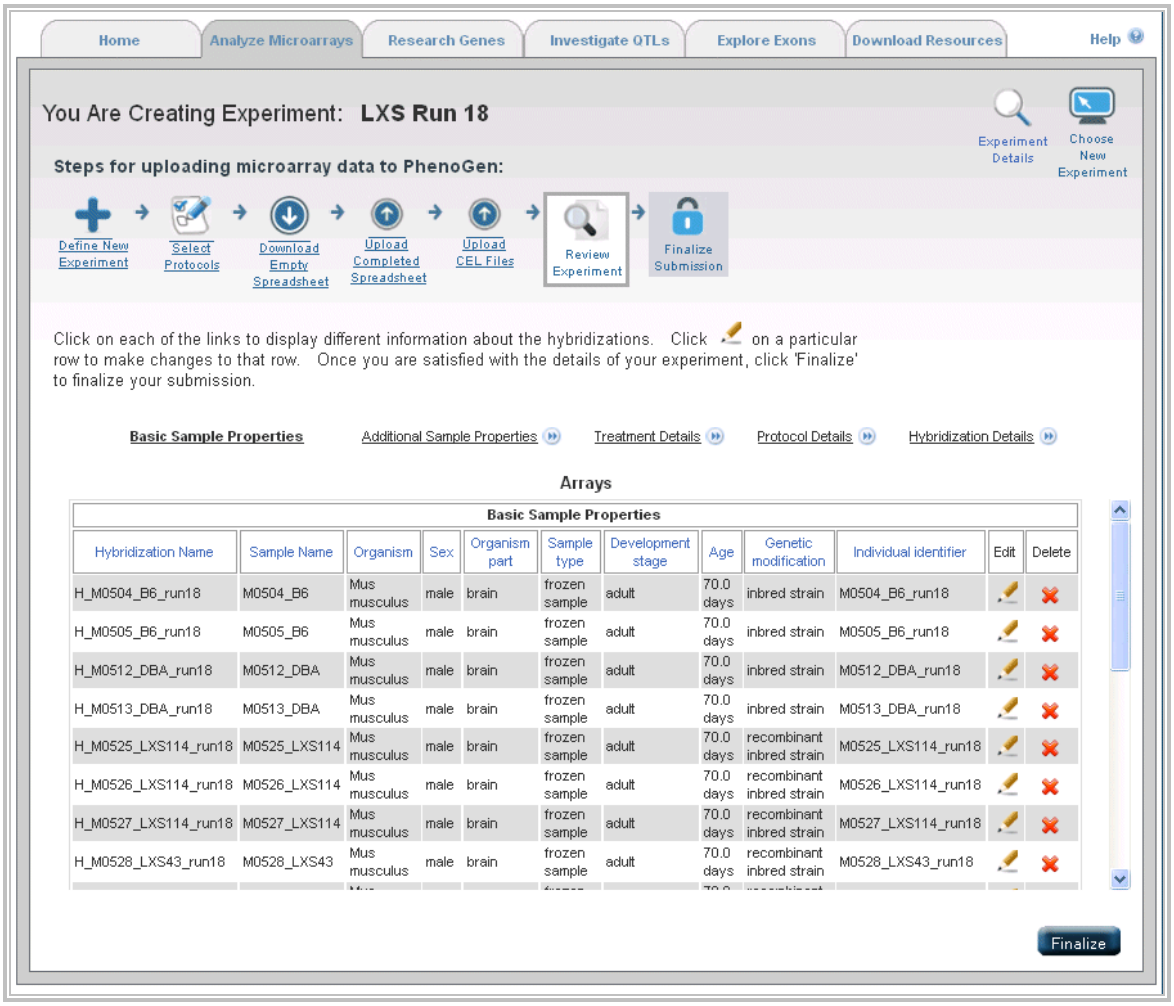
21. Click **Browse** beside each hybridization to find your data file (CEL or TXT format).
22. Click **Upload File(s)** when each hybridization has an associated array file.



23. Click **Next** after the files are uploaded.

Reviewing the Experiment

24. Click any of the links above the Arrays table to view details from that section of the spreadsheet. Your options are:
    - Basic Sample Details
    - Additional Sample Properties
    - Treatment Details
    - Protocol Details
    - Hybridization Details
25. Click the **pencil** icon to edit existing details for the displayed option (e.g., Basic Sample Properties). Click the red **X** to delete an array.



Finalizing the Submission

Important! After you finalize your submission, you cannot edit your experiment. See "Editing Your Experiments" for instructions on editing before finalization.

26. Click **Finalize** when you are certain your experiment details are correct.

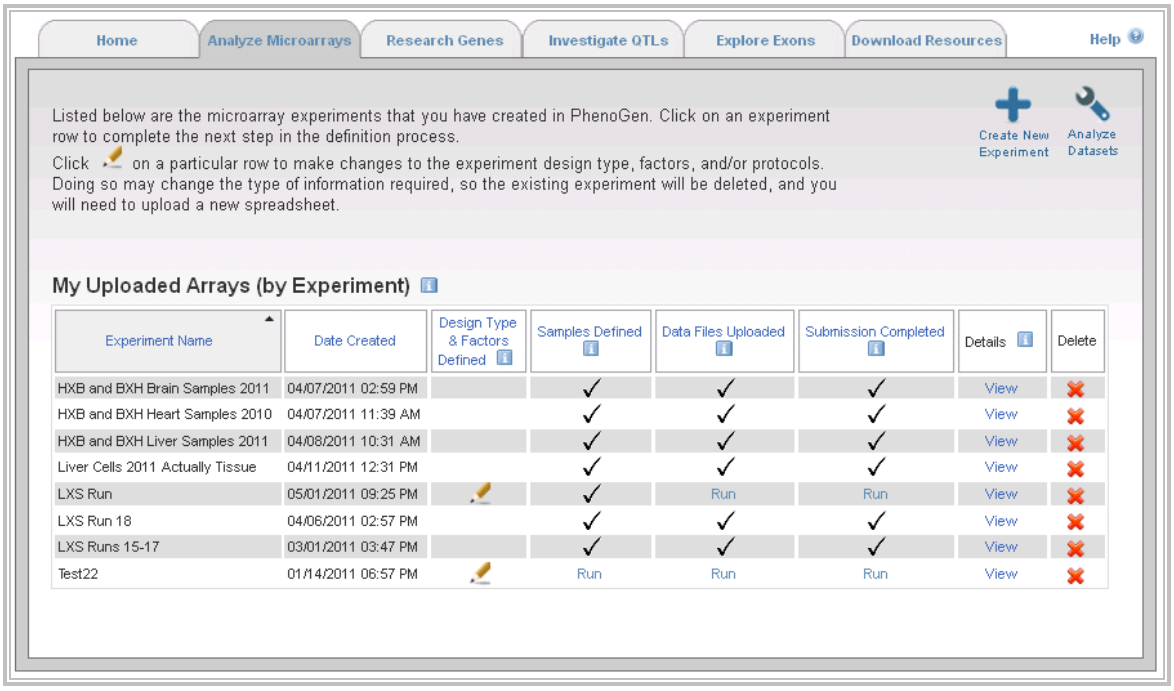
Editing Your Experiments

If you want to change the experiment design types or factors of an experiment, you must edit your experiment and download a new empty spreadsheet.

1. Click **Put my arrays onto this web site** in the *What would you like to do* section of the **Home** tab. The *Upload Arrays* page displays and shows a list of Microarray Experiments you have entered.

   OR

2. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
3. Click **Upload your own data** in the *What would you like to do* section. The *Upload Arrays* page displays and shows a list of Microarray Experiments you have entered.
4. Click the **pencil** icon beside the experiment you want to edit.



5. Proceed through the steps for "Creating an Experiment" on page 20.

Note: You must upload the completed spreadsheet again.

Creating Datasets

Creating datasets comprises four steps:

> "1. Retrieving Arrays" on page 25, to determine which arrays you want in the dataset.
>
> "2. & 3. Selecting and Finalizing Arrays" on page 27, to add them to the dataset you are creating and finalize the dataset with the selected arrays.
>
> "4. Running Quality Control" on page 28, on the dataset and reviewing the results.

1. Retrieving Arrays

Arrays in the PhenoGen website are uploaded into a local, MIAME-compliant database. There are public arrays which are available to any user and semi-public arrays which users can use after they are granted permission from the Principal Investigator responsible for the array.

Important! The PhenoGen website uses a local database for storing information about arrays, and data entered is only available on the PhenoGen website. See "MIAME Overview" for details about the Minimum Information About a Microarray Experiment (MIAME) standard.

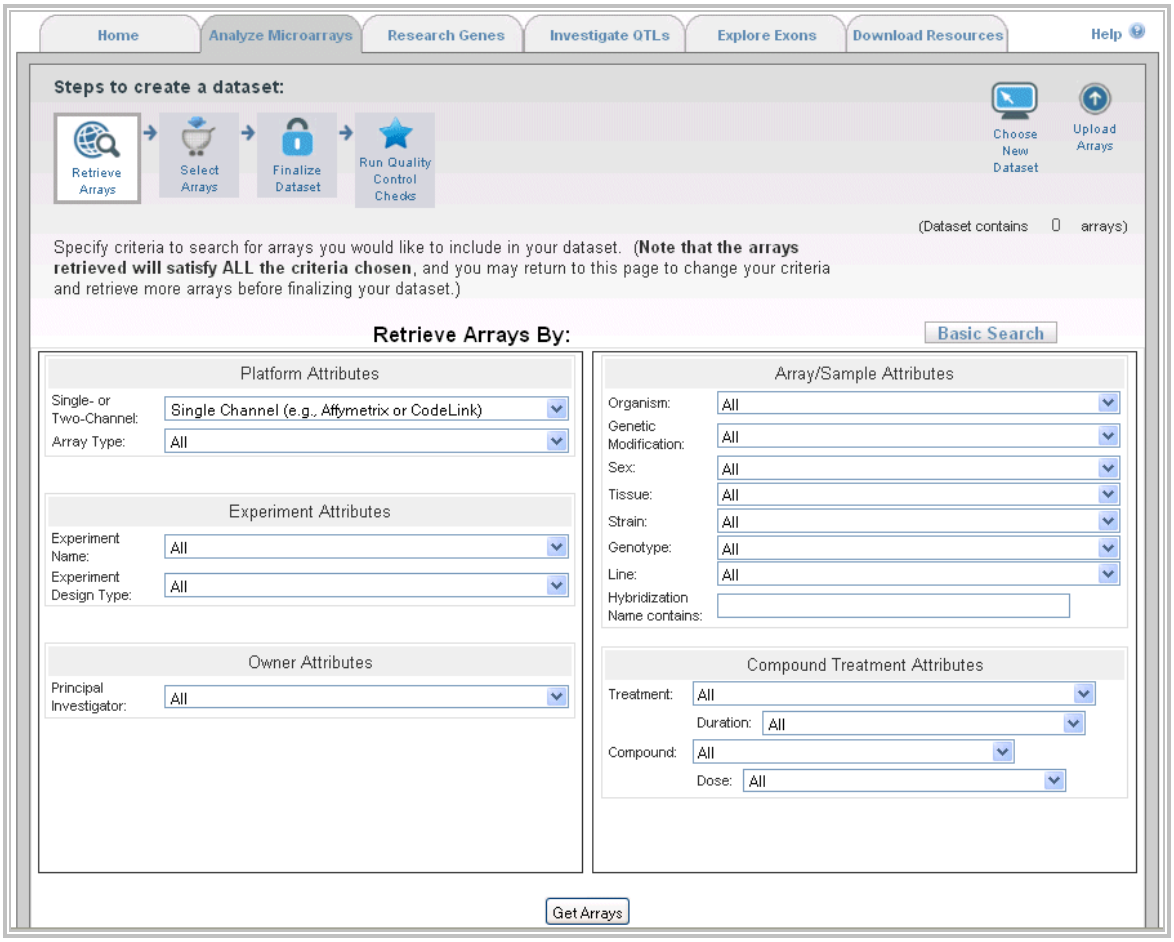
The PhenoGen website allows you to take single or multiple arrays from multiple lab experiments and combine them into "in-silico" datasets. Each array is annotated as displayed on the *Array Details* page. The annotation provides details such as species, gender, and array type. The set of arrays in the dataset can then be grouped by a particular set of characteristics (e.g., treated vs. untreated) and analyzed using the tools provided by the website. The goal of such an analysis is a list of genes that can be further investigated.

When retrieving arrays, you can filter a number of ways:

- Platform Attributes
  - Single or Two-Channel
  - Array type (e.g., Codelink_Rat_Whole_Genome)
- Experiment Attributes
  - Experiment name
  - Design type
- Owner Attributes
  - Principal Investigator
- Array/Sample Attributes
  - Organism
  - Genetic Modification (e.g., congenic strain)
  - Sex
  - Tissue (e.g., brain)
  - Strain (e.g., 1 HXB)
  - Genotype (e.g., hAC7 transgenic)
  - Line (e.g., alcohol accepting)
  - Hybridization Name contains…
- Compound Treatment Attributes
  - Treatment (e.g., control)
    - Duration
- Compound (e.g., saline)
  - Dose

Although filtering is not required, it is recommended.

The following image shows the Advanced Search to retrieve arrays. The Basic Search contains less filters.



Notes:

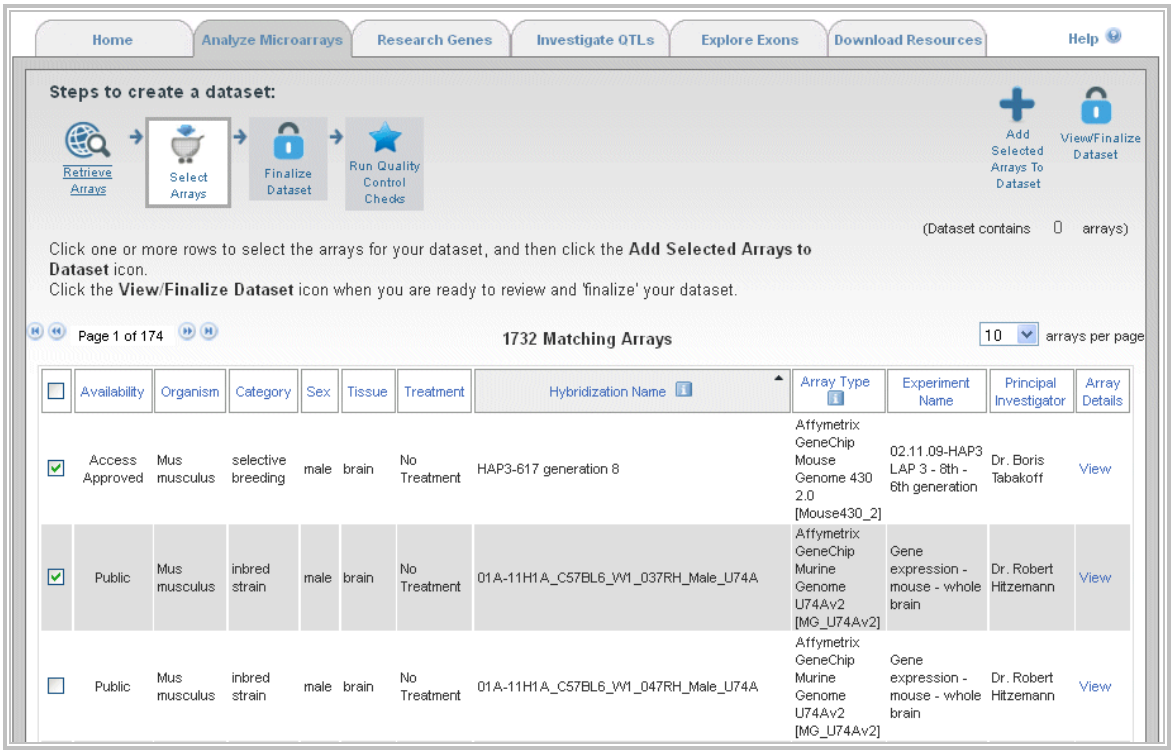
- Click **Advanced Search** to display more filtering options. Click **Basic Search** to display fewer options.
- Your choices in the drop-down lists are based on the experiments that have been uploaded into PhenoGen and your selection of Single Channel or Two Channel.
- Use the drop-down lists to narrow your criteria and limit the number of results that are returned.
- Leave the drop-down lists set to **All** to return all the arrays in the database. This is NOT recommended.
- Type a specific array name in the **Hybridization Name contains** field to retrieve only arrays that match your input.

2. & 3. Selecting and Finalizing Arrays

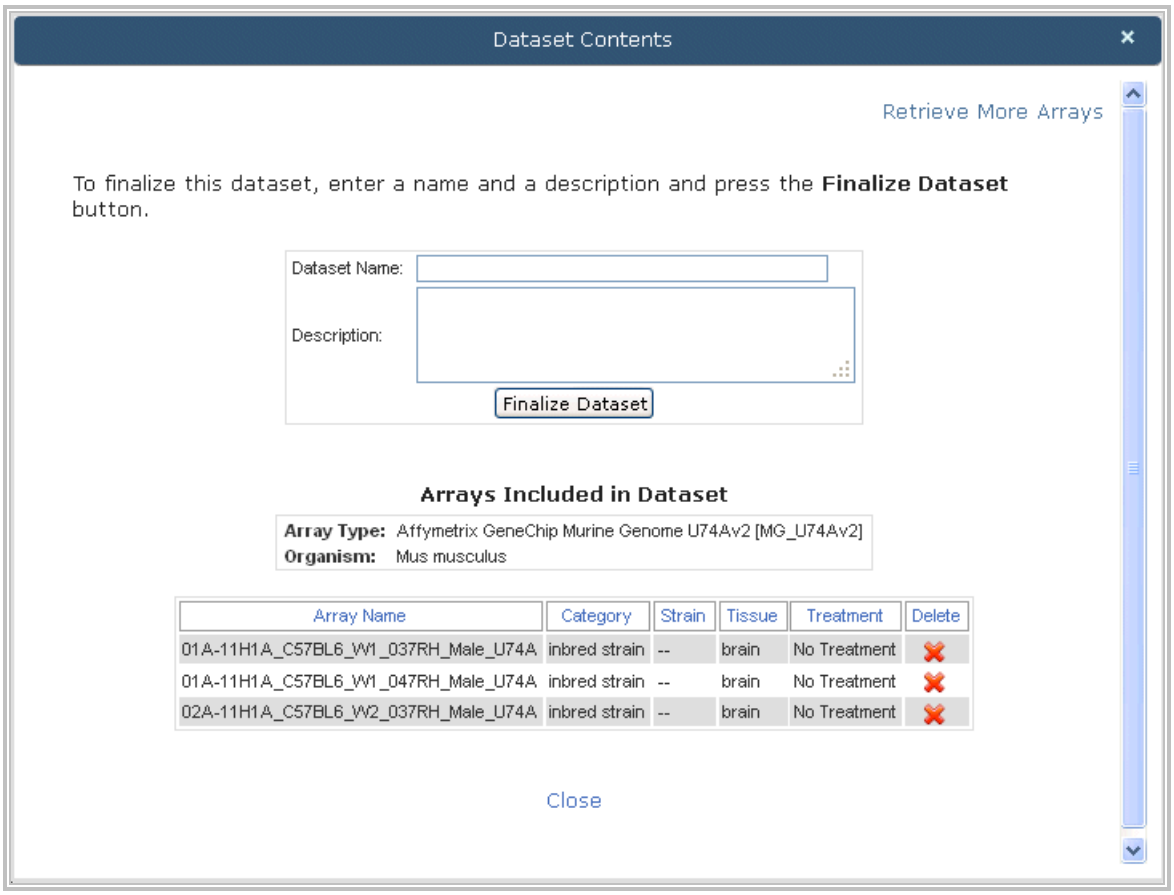
After you retrieve arrays, you can select the arrays that you want to be part of your dataset from the resultant list. If the **Availability** column shows "Access Required", the Principal Investigator who owns those arrays must give you permission to access them if you want to include them in your dataset. Open access arrays do not require permission to use and show "Public" in the **Availability** column.

Note: Array data is the responsibility of the Principal Investigator (PI). When you select arrays for which access is required, an email that requests access is sent to each of the Principal Investigators who are responsible for the data. When you are granted permission, you receive an email at the address you provided during registration.

If you want to review the array details, click **View** in the **Array Details** column to view the array information, or click the experiment name.



After you select and add the arrays you want to your dataset, click the **View/Finalize Dataset** link at the right to review and modify the arrays in your dataset, and finalize the dataset. If you require permission to use any of the arrays, your dataset remains in *Pending* status until permission is granted.



4. Running Quality Control

Quality control is an essential process when creating datasets. There are two quality control checks that ensure that the arrays you want to combine are compatible. They are:

- "Array Attribution Comparison (Step 1)" on page 35
- "Array Integrity (Step 2)" on page 35

When you run quality control, a quality control check of the selected arrays in the finalized dataset is performed. Arrays that are identified as questionable at any or all of the steps should be considered for deletion. However, some of the small imperfections and minor concerns can be alleviated by an appropriate normalization method. See "Preparing Datasets".

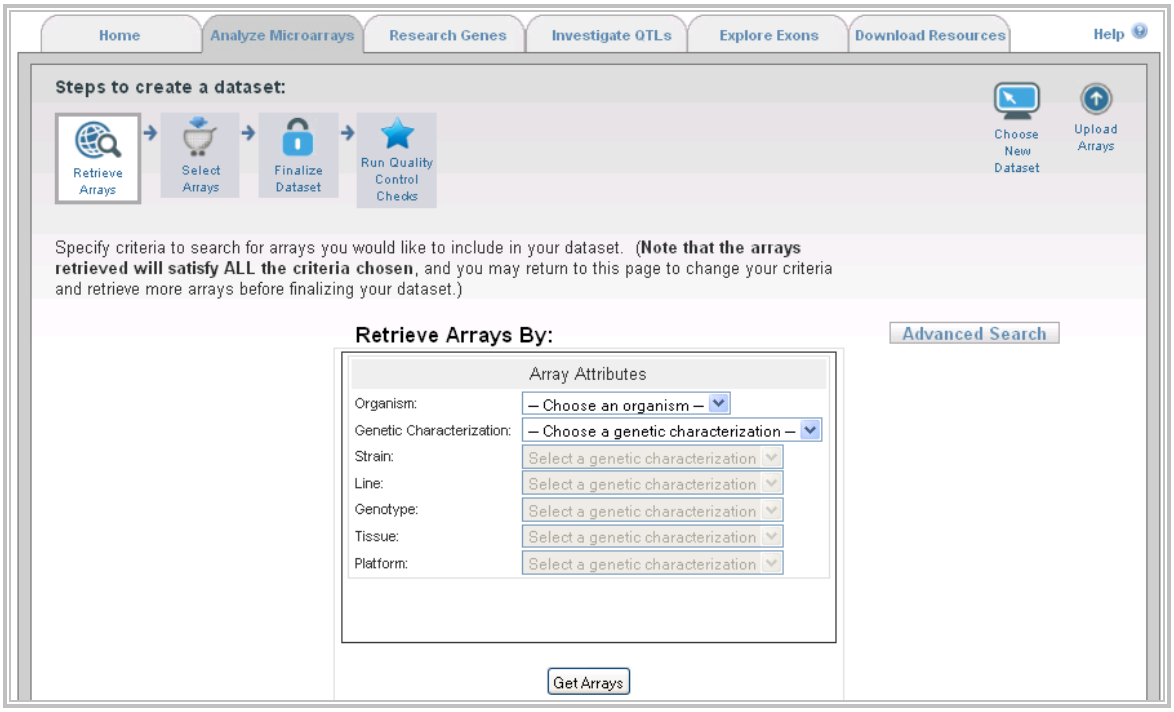
For more details about the QC procedures commonly used for microarrays, see "Additional Quality Control Sources" on page 154.

Retrieving Arrays

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click the **Create a dataset** link in the *What would you like to do* section. The Retrieve Arrays page displays.

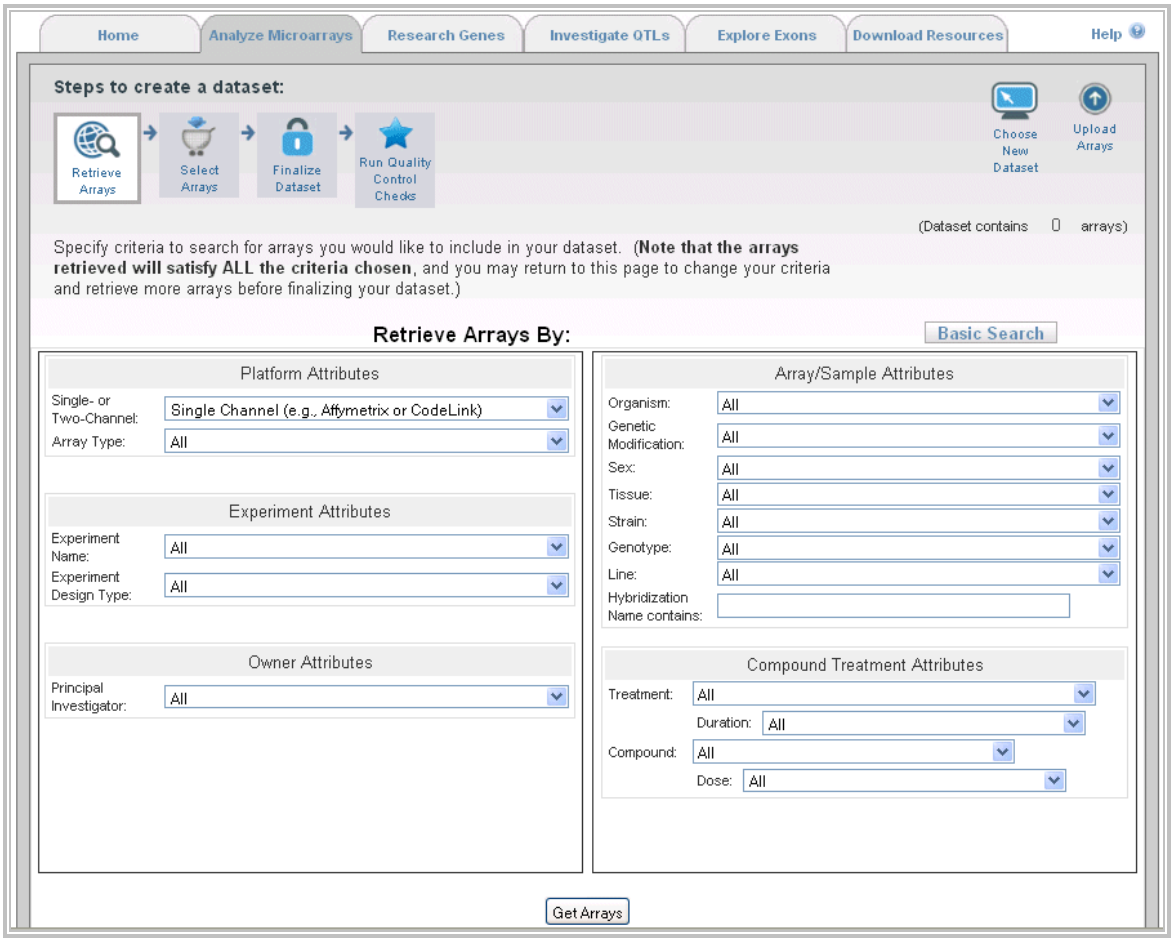
Basic Search:

3. Choose an **Organism** and a **Genetic Characteristic** from the drop-down lists. The choices you make determine which other fields are available.
4. Choose a strain, line, genotype, tissue, and platform from the available options.
5. Click **Get Arrays**. Arrays that match all of the specified criteria display.



Advanced Search:

3. Click **Advanced Search** if you would like to choose more filtering criteria. Or, choose options from the Basic Search.



4. Choose options from the drop-down lists in the *Platform Attributes* section.
5. Choose options from the drop-down lists in the *Experiment Attributes* section.
6. Choose options from the drop-down lists in the *Owner Attributes* section.
7. Choose options from the drop-down lists in the *Array/Sample Attributes* section.
8. Enter the whole or partial array name in the **Hybridization Name Contains** field.
9. Choose options form the drop-down lists in the *Compound Treatment Attributes* section.
10. Click **Get Arrays**. Arrays that match all of the specified criteria display.

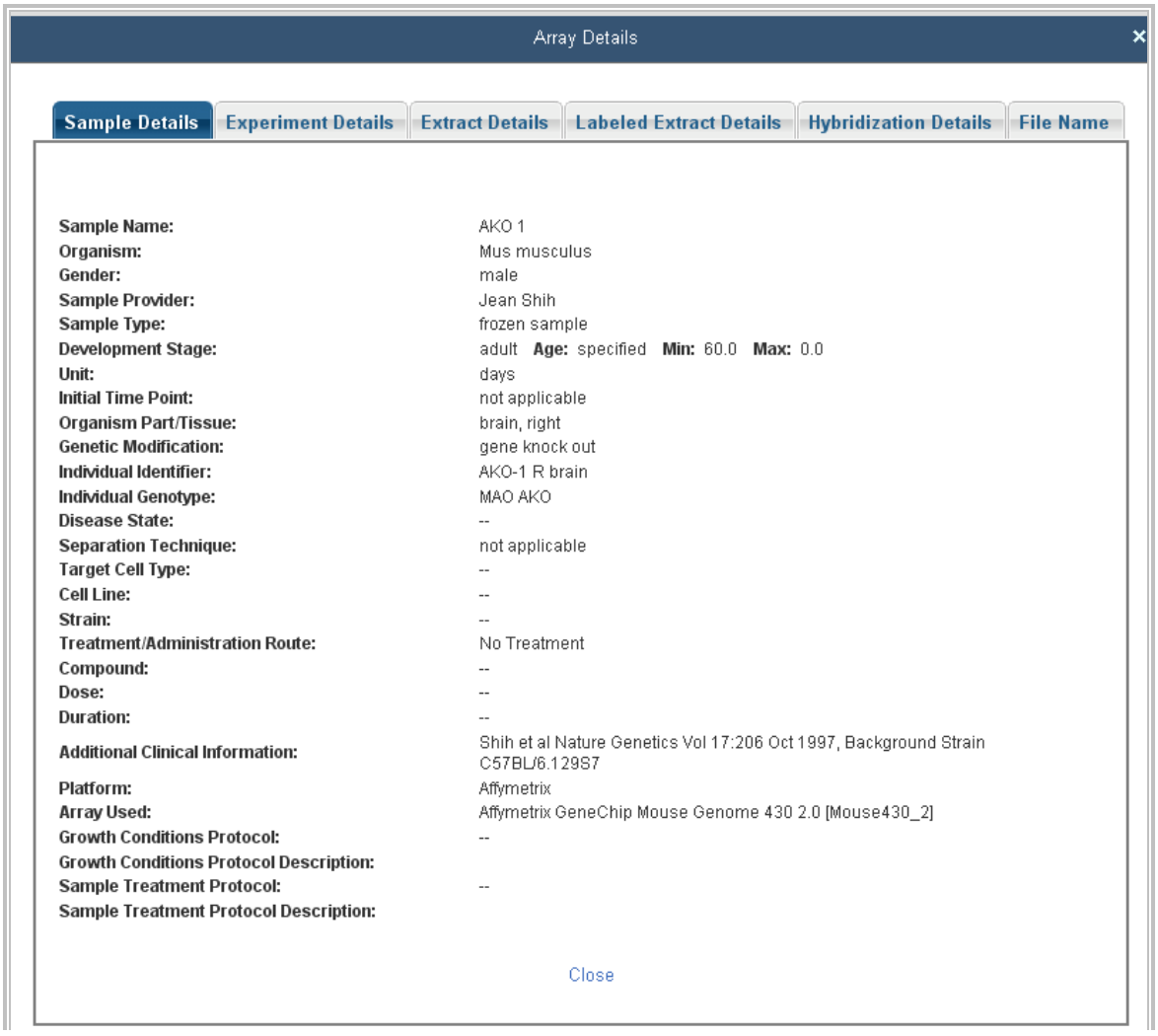
Note: You can repeat the preceding steps as many times as necessary to retrieve arrays that match various criteria.

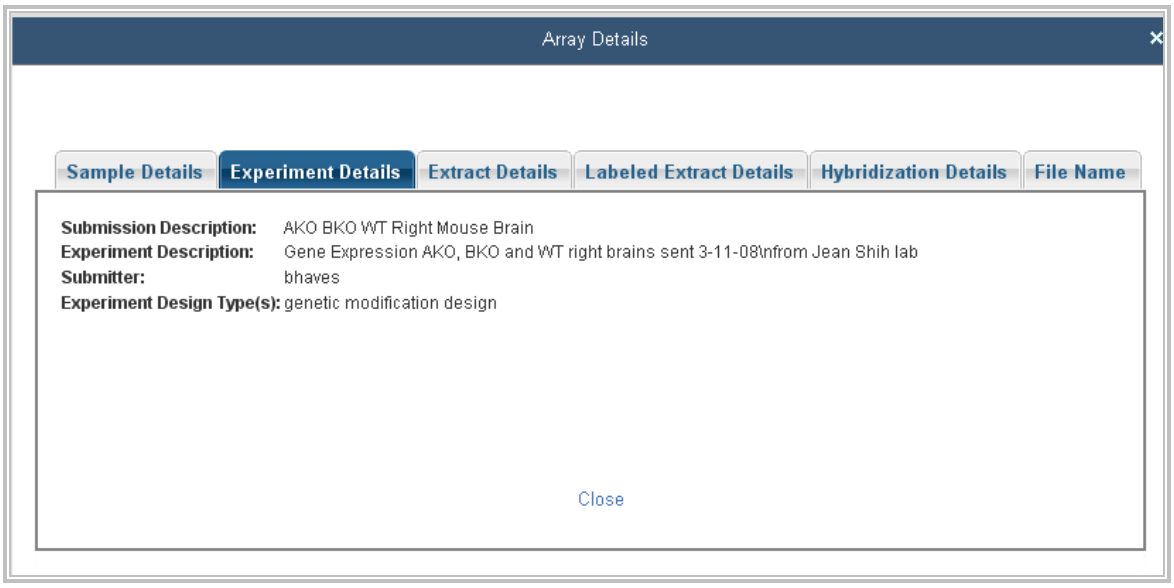
Viewing Array Details

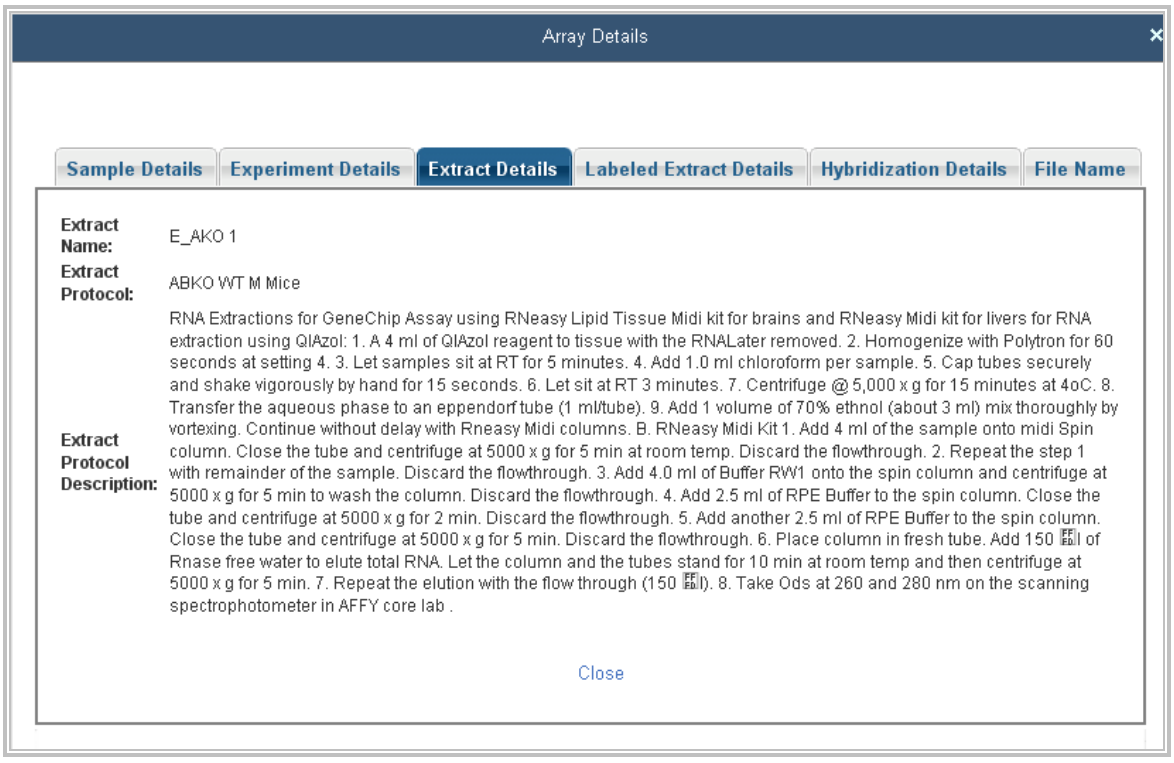
You can review details about the arrays on the *Array Details* page to determine if you want to select them for your dataset.

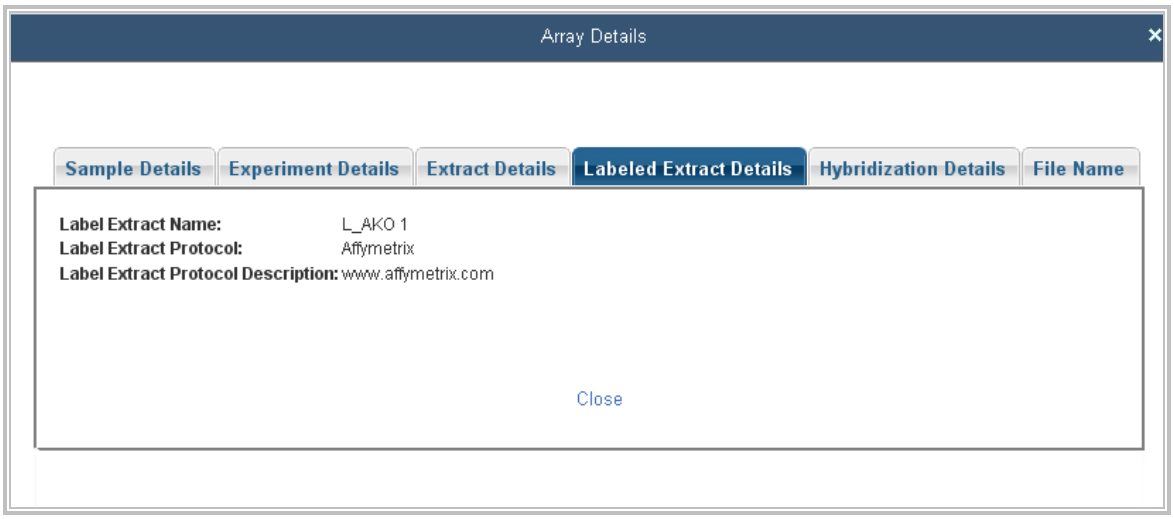
1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click the **Create a dataset** link.
3. Retrieve arrays. See "Retrieving Arrays".
4. Click **View** in the **Array Details** column to view details about the sample and array.

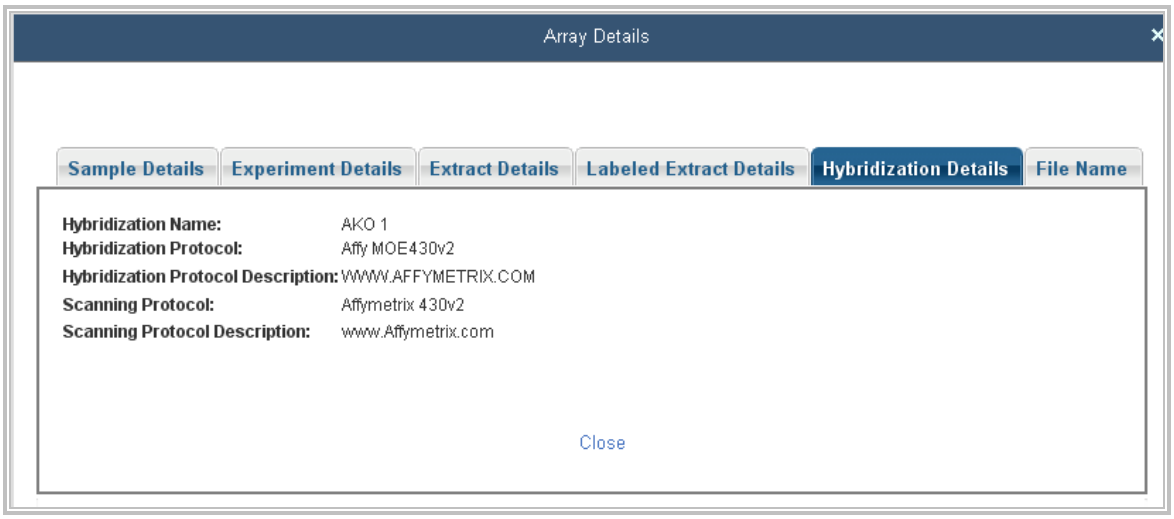
Array Details are shown on six tabs: Sample Details, Experiment Details, Extract Details, Labeled Extract Details, Hybridization Details, and File Name. Click the tab that contains the information that you want to view.




Array Details
Sample Details
Experiment Details
Extract Details
Labeled Extract Details
Hybridization Details
File Name
Submission Description: AKO BKO WT Right Mouse Brain
Experiment Description: Gene Expression AKO, BKO and WT right brains sent 3-11-08\nfrom Jean Shih lab
Submitter: bhaves
Experiment Design Type(s): genetic modification design
Close


Array Details
Sample Details
Experiment Details
Extract Details
Labeled Extract Details
Hybridization Details
File Name
Extract Name: E_AKO 1
Extract Protocol: ABKO WT M Mice
Extract Protocol Description: RNA Extractions for GeneChip Assay using RNeasy Lipid Tissue Midi kit for brains and RNeasy Midi kit for livers for RNA extraction using QIAzol: 1. A 4 ml of QIAzol reagent to tissue with the RNALater removed. 2. Homogenize with Polytron for 60 seconds at setting 4. 3. Let samples sit at RT for 5 minutes. 4. Add 1.0 ml chloroform per sample. 5. Cap tubes securely and shake vigorously by hand for 15 seconds. 6. Let sit at RT 3 minutes. 7. Centrifuge @ 5,000 x g for 15 minutes at 4oC. 8. Transfer the aqueous phase to an eppendorf tube (1 ml/tube). 9. Add 1 volume of 70% ethnol (about 3 ml) mix thoroughly by vortexing. Continue without delay with Rneasy Midi columns. B. RNeasy Midi Kit 1. Add 4 ml of the sample onto midi Spin column. Close the tube and centrifuge at 5000 x g for 5 min at room temp. Discard the flowthrough. 2. Repeat the step 1 with remainder of the sample. Discard the flowthrough. 3. Add 4.0 ml of Buffer RW1 onto the spin column and centrifuge at 5000 x g for 5 min to wash the column. Discard the flowthrough. 4. Add 2.5 ml of RPE Buffer to the spin column. Close the tube and centrifuge at 5000 x g for 2 min. Discard the flowthrough. 5. Add another 2.5 ml of RPE Buffer to the spin column. Close the tube and centrifuge at 5000 x g for 5 min. Discard the flowthrough. 6. Place column in fresh tube. Add 150 µl of Rnase free water to elute total RNA. Let the column and the tubes stand for 10 min at room temp and then centrifuge at 5000 x g for 5 min. 7. Repeat the elution with the flow through (150 µl). 8. Take Ods at 260 and 280 nm on the scanning spectrophotometer in AFFY core lab .
Close


Array Details
Sample Details
Experiment Details
Extract Details
Labeled Extract Details
Hybridization Details
File Name
Label Extract Name: L_AKO 1
Label Extract Protocol: Affymetrix
Label Extract Protocol Description: www.affymetrix.com
Close


Array Details
Sample Details
Experiment Details
Extract Details
Labeled Extract Details
Hybridization Details
File Name
Hybridization Name: AKO 1
Hybridization Protocol: Affy MOE430v2
Hybridization Protocol Description: WWW.AFFYMETRIX.COM
Scanning Protocol: Affymetrix 430v2
Scanning Protocol Description: www.Affymetrix.com
Close


Array Details
Sample Details
Experiment Details
Extract Details
Labeled Extract Details
Hybridization Details
File Name
File Name: hopkins_AKO1_RB_Mouse430_2.CEL
Close

Selecting Arrays & Finalizing Datasets

After you retrieve arrays, you can select the ones you want to use in your dataset, then finalize the dataset.

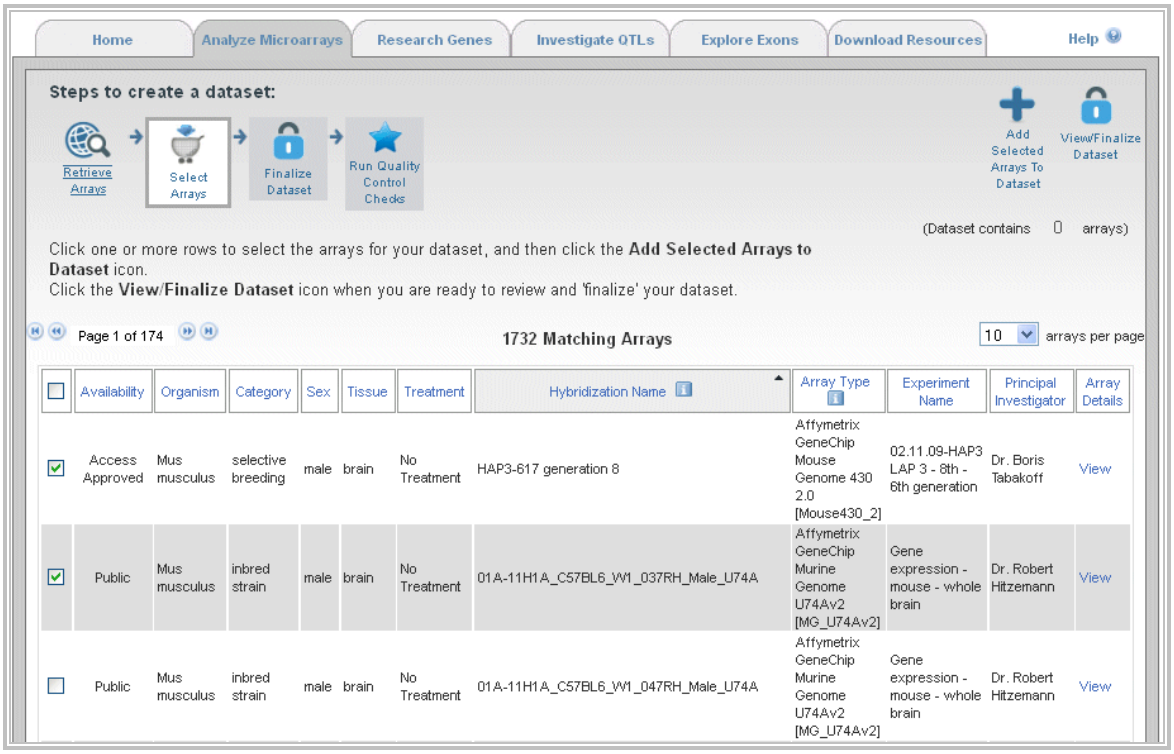
Selecting Arrays for a Dataset

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click the **Create a dataset** link in the *What would you like to do* section. The Retrieve Arrays page displays.
3. Retrieve arrays. See "Retrieving Arrays".

Notes:

- Use the drop-down list in the **Display [number] arrays per page** field to choose the number of arrays to display.
- Click any blue column heading to sort the arrays by that column.

4. Click the arrays or the checkbox beside the arrays you want to add to your dataset. Click the checkbox in the column header to select all arrays.



Notes:

- The **Availability** column shows array accessibility:
  - Open access data shows *Public.*
  - Semi-public data shows *Access Required*.
  - If you requested access that has not yet been granted, the column shows *Access Pending*.
  - When you receive permission to access an array, the column shows *Access Approved*.
  - If you are denied access, the column shows *Access Denied*.

5. **OPTIONAL:** Click **View** in the **Array Details** column to view sample information for an array.
6. Click the **Add Selected Arrays to Dataset** link when you have selected the arrays you want. A confirmation message displays. Click **Close**.
7. Click **View/Finalize Dataset** link to view the dataset and the selected arrays. The *Finalize Dataset* page displays.

Finalizing a Dataset

After you select the arrays you want to use in your dataset, you can finalize the dataset.

1. Enter a **Dataset Name** and **Description**.
2. Click **Finalize Dataset**. You dataset is finalized, and a confirmation message displays. Click **Close**.

Your dataset displays in the list of datasets, where you can run quality control, group and normalize, and save the resultant gene list. If access to arrays is pending, the dataset shows *Pending*.

Quality Control Checks Overview

The PhenoGen website runs two quality control checks to ensure that the arrays you want to combine are compatible:

- Array Attribution Comparison
- Array Integrity

For more details about the QC procedures for commonly used microarrays, see "Additional Quality Control Sources" on page 154.

Array Attribution Comparison (Step 1)

The information for the arrays is compared, and discrepancies are listed for the user. A table displays. Attributes that differ within a category are highlighted in orange text:

- Sex
- Sample type
- Development Stage
- Age
- Initial Time Point
- Organism Part
- Genetic Modification
- Individual Identifier
- Individual genetic trait or genotype
- Disease state
- Separation technique
- Cell type or target cell type
- Cell line
- Strain

Array Integrity (Step 2)

The quality control process looks specifically at each array. There are two steps:

1. Each array image is individually checked using measurements outlined by Affymetrix or CodeLink. See"Guidelines for Assessing Affymetrix Data Quality" on page 36 and "Guidelines for Assessing CodeLink Data Quality" on page 50.
2. Arrays within a dataset are compared. See "Within-Array Checks for Affymetrix 3' Arrays" and "Within-Array Checks for Affymetrix Exon Arrays" on page 41.

Note: Neither step indicates definitively whether an array is "bad". Instead, you must balance considerations for quality of data and quantity of data with respect to the analysis at hand.

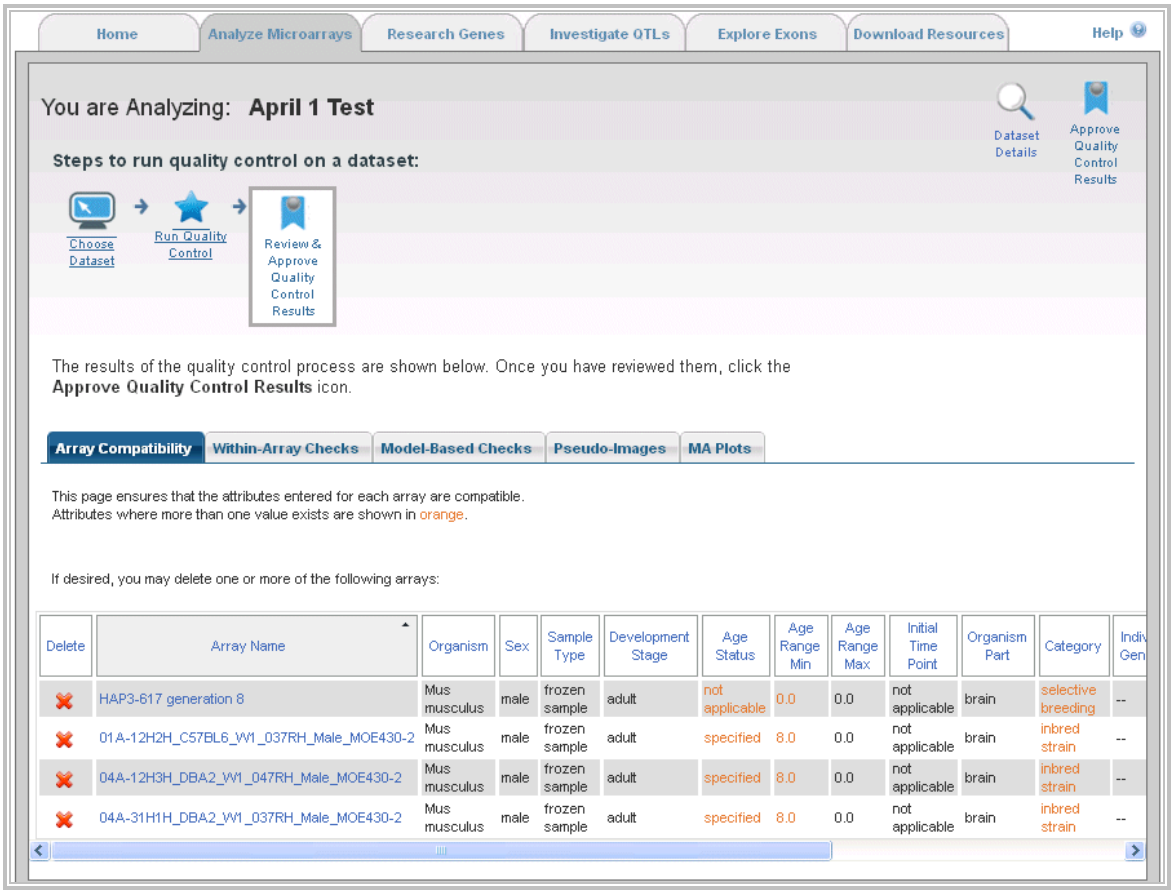
The output of this quality control check can be seen on tabs that displays graphs for determining whether the arrays are ready for analysis.

Guidelines for Assessing Affymetrix Data Quality

After you run quality control on Affymetrix datasets, graphs and tables display on individual tabs.

Notes:

- If you choose not to generate images when you run the quality control checks, the Pseudo Images and MA Plots tabs have no images.
- Click the Download icon to download the images from each tab.



See the following for explanations of the data that displays on each tab:

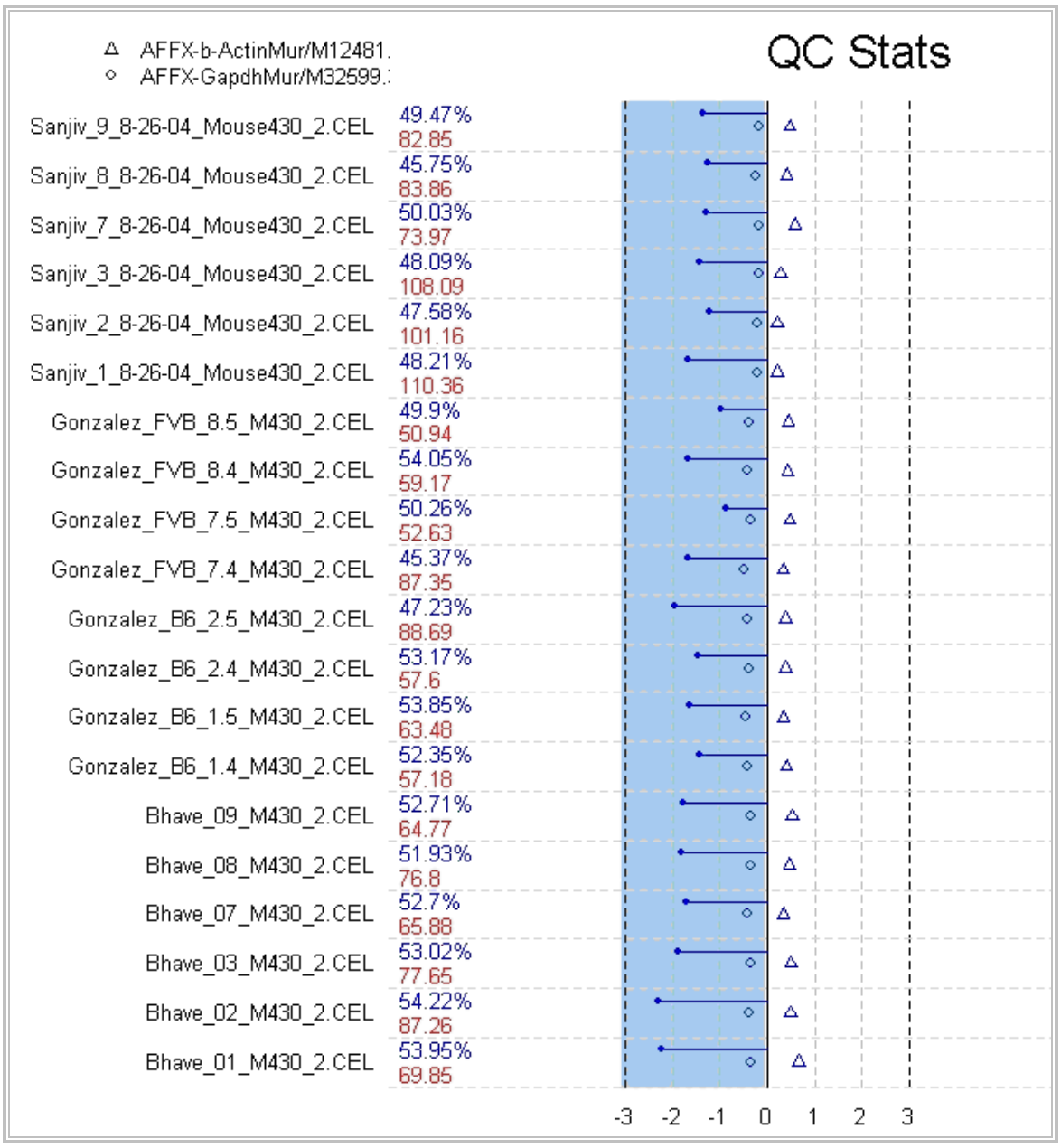
- "Within-Array Checks for Affymetrix 3' Arrays" on page 37.
- "Within-Array Checks for Affymetrix Exon Arrays" on page 41.
- "Model-based Checks for Affymetrix 3' Arrays" on page 39.
- "Model-based Checks for Affymetrix Exon Arrays" on page 43.
- "Pseudo Images (Affymetrix)" on page 46.
- "MA Plots" on page 48.

Within-Array Checks for Affymetrix 3' Arrays

The within-array quality control checks are examined using the Bioconductor package *Simpleaffy*. There are four checks that are examined:

- **Average Background** Average background is examined to determine if it is consistent across arrays. Affymetrix has indicated that typical background averages range from 20 to 100, but there is no statistically relevant range for these values to fall within.
- **Internal Controls** There are two internal house-keeping genes (β-actin and GAPDH) that are used to evaluate the RNA and assay quality. Three probe sets have been designed per control. The first probe set measures the intensity of the 3' end of the gene, the second probe set measures the intensity of the 5' end of the gene, and the third probe set measures the intensity in the middle of the gene. The ratio of the intensity from the 3' end to the 5' end should theoretically be around 1. According to Wilson, et al. (2004), ratios above 1.25 for GAPDH should be considered outliers and ratios over 3 for β-actin should be considered outliers.
- **Percent Present** Affymetrix recommends the use of a normalization and summary method called Microarray Suite 5.0 (MAS5). Within this normalization procedure, each probe set gives a Present, Marginal, or Absent call. The percent of present probe sets out of all probe sets on the array is used as a quality control measure. Although the percent of present probe sets measured is highly dependent on each specific experiment with respect to the number of genes you expect to be expressed, an extremely low value raises suspicion about the quality of an array. Also, it is expected that duplicate arrays have similar percent missing levels.
- **Scaling Factors** As part of the quality control procedure, intensities are normalized using the MAS5 procedure. Within the process of normalization, each array is adjusted by a scaling factor to get the trimmed mean of all arrays to equal a target signal. This scaling factor indicates how much RNA was hybridized onto the array. A wide variation of scaling factors across arrays can be a cause for concern. Affymetrix defines a wide variation as a three-fold or greater difference.

The *Simpleaffy* package from Bioconductor calculates the four within-array quality control check measures from a group of CEL files and displays the results on a single QC Stats graph:



Along the left side of the graph are the names of the CEL files that were included in this analysis. The next column has two numbers per CEL file. The top number is the *percent present* and the bottom number is the *average background* for that CEL file. It is expected that the *average background* measures across arrays should be similar and, ideally, below 100.

The *percent present* values are heavily dependent on the type of sample used on the array. If the same type of tissue is used in all samples, the percent present values should be similar across arrays. However, if you have multiple tissue types such as liver and brain, the percent present values could vary substantially between these tissues.

When the *average background* measures display in red, it indicates that the values across arrays show a "considerable amount of variation". When the *percent present* values display in red, it indicates that there is a spread greater than 10% between the lowest and highest percentage.

The solid dots that are attached to a horizontal bar originating from the zero line represent the *scale factors* (indicators of how much RNA was hybridized to the array) for each array. The blue shading is the region that spans three-fold below and three-fold above the average *scale factor*. In the graph above, all of the *scale factors* fall into this range. However, if one scale factor did not fall within the range, the dot and horizontal line for that scale factor would display in red.

The intensity of the two *internal control* housekeeping genes is represented by open triangles and open diamonds and measures the quality of the hybridized RNA.

The open triangles represent the log base 2 of the 3' to 5' ratio for *β-actin*. In the graph above, a value of 0 for the ratio would be ideal and a value above 1.6 would be a cause for concern. None of the ratios in the graph are greater than 1.6.

Similarly, the open diamonds on the graph represent the log base 2 of the 3' to 5' ratio of *GAPDH*. This ratio should be below 0.32. Again, none of the ratios are above the threshold.

Model-based Checks for Affymetrix 3' Arrays

Another package in Bioconductor looks at a model-based quality control assessment. There are three assessments that are examined at this stage; relative log expressions (RLE), normalized unscaled standard errors (NUSE), and array pseudo-images. To calculate all three assessments, a probe level model must first be fit to the data:

$$\log_2 PM_{kij} = \beta_{kj} + \alpha_{ki} + \varepsilon_{kij}$$

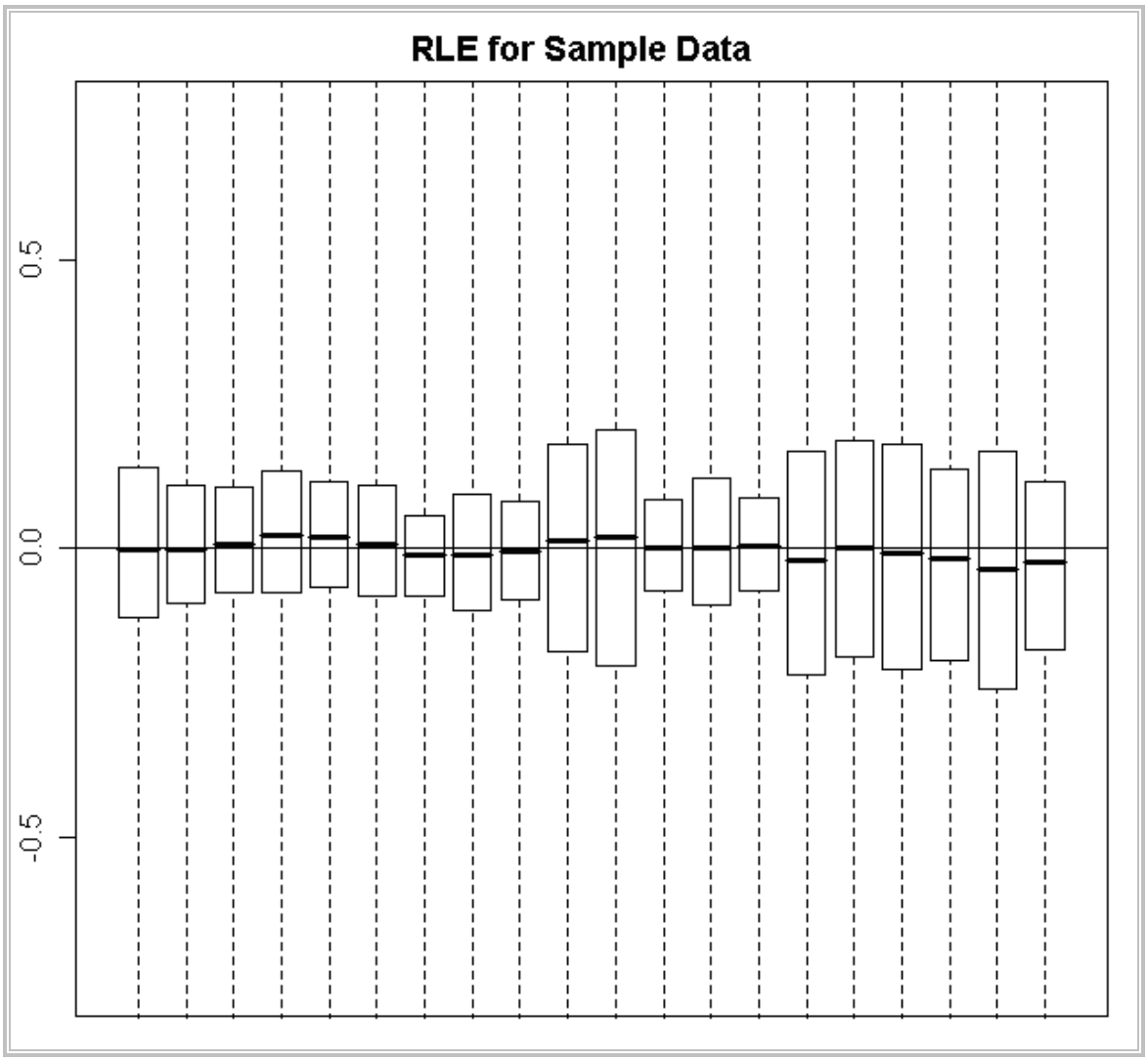
k = 1 ... K probe sets

i = 1 ... I_k probes

j = 1 ... J arrays

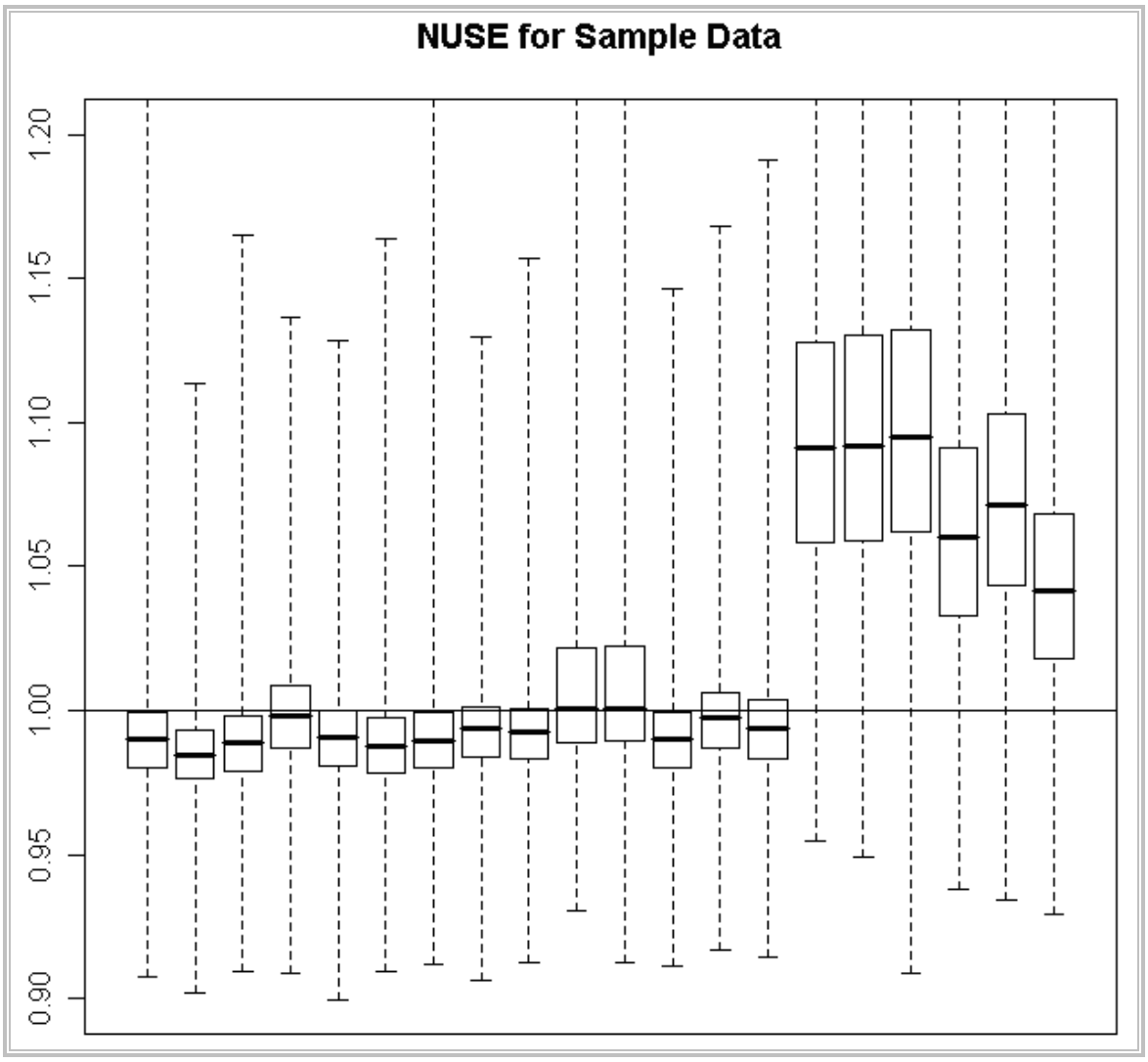
In this default model, β_{kj} is the array effect, α_{ki} is the probe effect, and ε_{kij} is the residual error term. The model can be adjusted to include other effects, but the default model is used for quality control purposes.

Relative log expressions (RLE) The relative log expressions for each probe represent that particular probe's deviation from the median value of that probe across arrays. This quality assessment is dependent on the assumption that most of the genes measured are expressed at similar levels across the arrays. The relative logs are displayed as box plots. The expectation is that the relative log expressions should be evenly distributed around zero within each array, i.e., one array does not always have a higher intensity than all the other arrays when looking at individual probes. Also, if one or more arrays have box plots that are much larger than the other arrays, then these arrays tend to have more outliers than the other arrays.



The RLE graph above displays a variety of box lengths and several boxes that are not centered around 0. This raises concerns about the distribution of the RLEs within arrays, but these minor issues could be resolved with normalization.

Normalized Unscaled Standard Errors (NUSE) The normalized unscaled standard errors represent the standard error between probe intensities within a probe set on a specific array. These errors are normalized by dividing by the median standard error for that probe set across arrays. The expected distribution of NUSEs within an array is centered about one. A higher value indicates that the array has more variance for that probe set than the other arrays.



The plot above is a concern because the six samples on the right side have much more variation than the other 14 samples. In other words, the variation between probes within a probeset is consistently higher in these six samples. The extent of the variation indicates that the samples could be of poorer quality.

Within-Array Checks for Affymetrix Exon Arrays

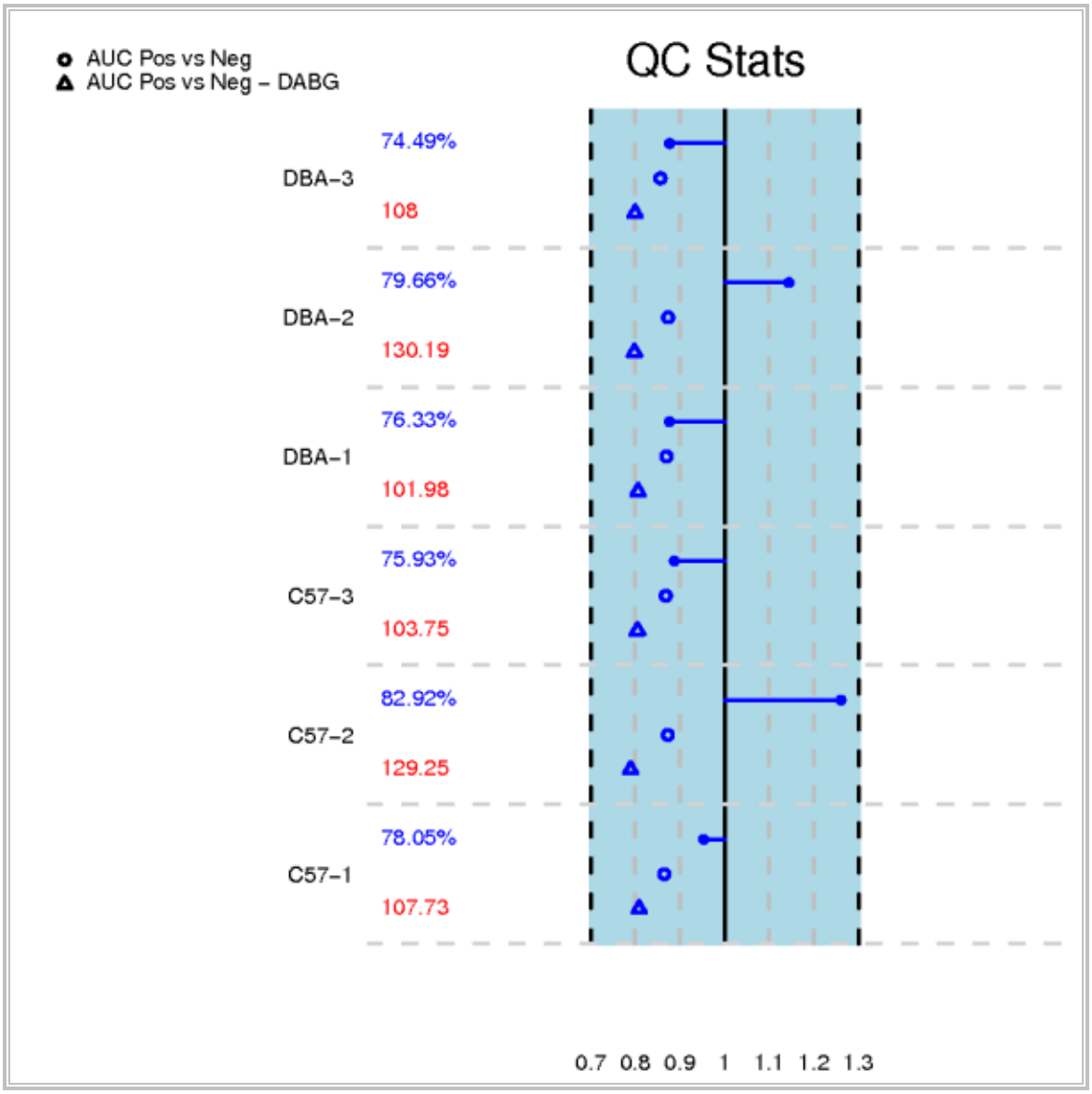
The within-array quality control checks are examined using data from quality control reports produced by the Affymetrix Power Tools and graphics from the Bioconductor package *Simpleaffy*. Quality control measures are based on an RMA normalization of the core transcript clusters. There are four checks that are examined:

- **Average Background** The average raw intensity value across probes used to calculated background. The average is calculated on intensity values prior to any normalization. There is no relevant threshold for this value. Look for consistency across arrays.
- **Percent Present** The proportion of all probe sets (transcript clusters) with an intensity value above detection limits (p-value < 0.01). Although the percent of transcript clusters above detection limits is highly dependent on each specific experiment with respect to the number of genes you expected to be expressed, an extremely low value raises suspicion about the quality of an array. It is expected that duplicate arrays have similar percent-missing levels.
- **Pseudo Scaling Factors** Represents the replicated scaling factor from the quality control measure for the 3' expression arrays. It is the ratio of the average raw intensity value of all probes on the array compared to the average raw intensity value across all arrays in the experiment. This value gives a

general idea of how ‘dim’ or ‘bright’ an array is. Most minor discrepancies among arrays are eliminated with proper normalization.

- **AUC for Distinguishing Positive and Negative Controls** The positive and negative controls are used in a receiver operating characteristic (ROC) analysis to assess the array's ability to distinguish between the two, based on signal intensity. The area under the curve (AUC) is a descriptive measure to assess this accuracy. An AUC of 1 indicates the perfect separation of positive and negative controls based on signal intensity, while an AUC of 0.5 indicates that signal intensity cannot be used to distinguish between the two types of probes. The AUC can also be calculated based on detection above background measures.

These measures are displayed in a graphic similar to the one generated by the *Simpleaffy* package from Bioconducter for the 3’ Affymetrix arrays.



Along the left side of the graph are the names of the CEL files that are included in the analysis. The next column has two numbers per CEL file: the top number is the *percent present* and the bottom number is the *average background* for that CEL file.

The percent present values are heavily dependent on the type of sample used on the array. If the same type of tissue is used in all samples, the percent present values should be similar across arrays. However, if you have multiple tissue types such as liver and brain, the percent present values can vary substantially between these tissues. By default, the percent present values are displayed in red if there is a spread greater than 10% between the lowest and highest values within the experiment. It is expected that the average background values across arrays should be similar. If they vary by more than 20 units, all values are given in red.

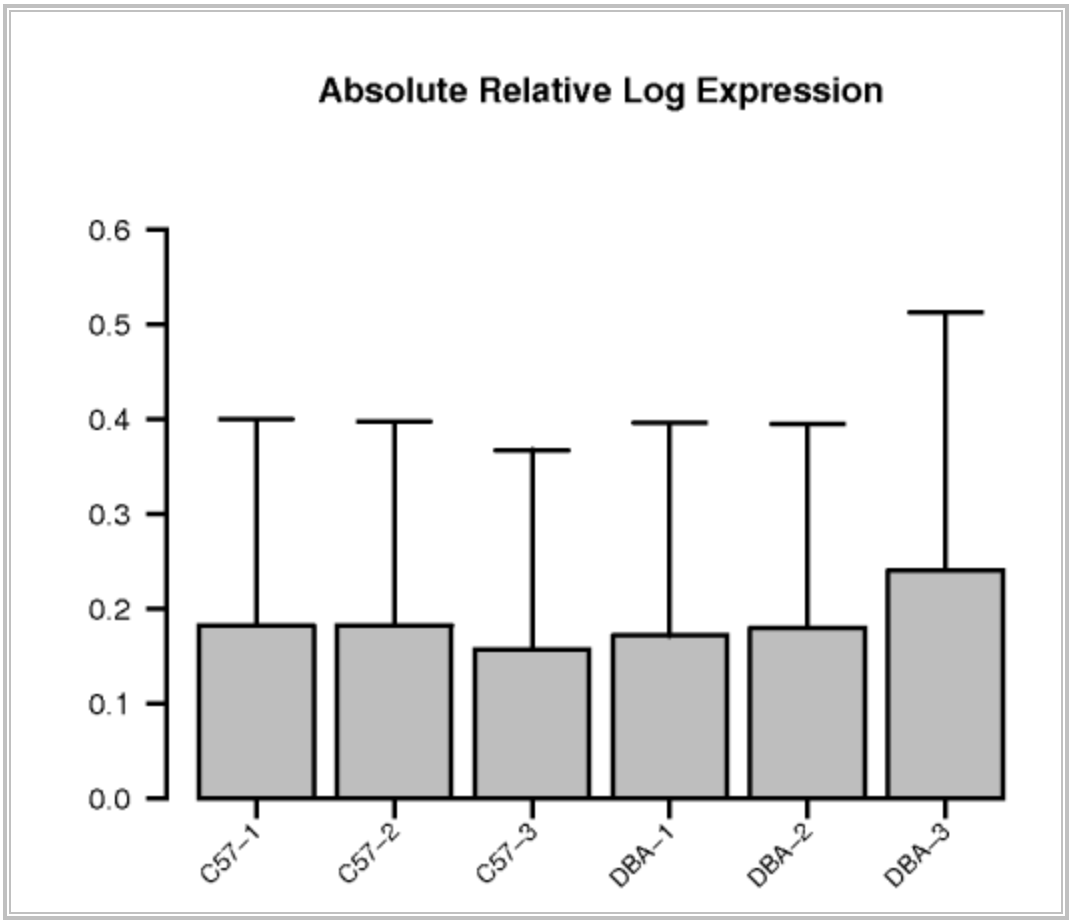
The solid dots that are attached to a horizontal line originating from the solid vertical line through 1 represent the pseudo-scaling factors (indicators of how much RNA was hybridized to the array) for each array. The blue shaded region is the area 30% above and 30% below the experiment average. Values that fall outside this region are displayed in red.

The open circles and open triangles represent the AUC for distinguishing positive and negative controls respectively, based on intensity values and detection above background values. According to Affymetrix, "values between 0.80 and 0.90 are typical" (*Quality Assessment of Exon and Gene Arrays, 2007*) for AUCs based on signal intensity. AUC values based on detection above background tend to have values between 0.75 and 0.85.

Model-based Checks for Affymetrix Exon Arrays

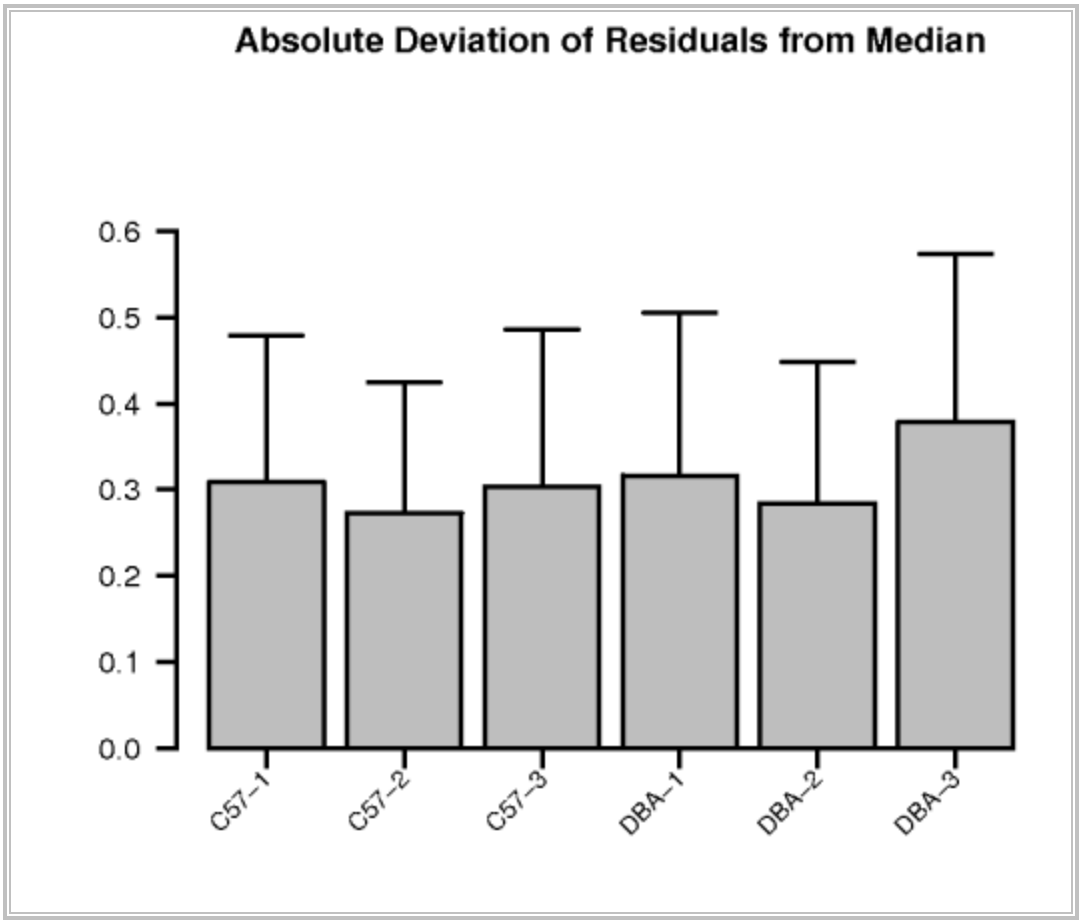
For the Affymetrix Exon Array, two model-based measures are explored that are similar to the model-based measures calculated for the Affymetrix 3' Arrays. These are relative log expression (RLE) and mean of the absolute deviation of the residuals (MAD). All measures are gathered from the summary report generated by the Affymetrix Power Tools when the experiment is normalized using RMA on the core transcripts.

Absolute Relative Log Expression (RLE). The absolute relative log expressions for each transcript cluster represent that particular transcript cluster's absolute deviation from the median value of that transcript cluster across arrays. Displayed in the RLE figure for exon arrays is the mean absolute relative log expression across transcript clusters within an array and the standard deviation of this value within an array. Consistent values across arrays are ideal. If an array has a higher mean or a significantly larger standard deviation, the quality of this array may be suspect.



The RLE graph above displays consistent results across samples. The final sample, DBA_3, appears to have a higher mean RLE, indicating that the intensities for that particular array deviated to a greater extent from the other arrays. However, small deviation is considered only a minor issue.

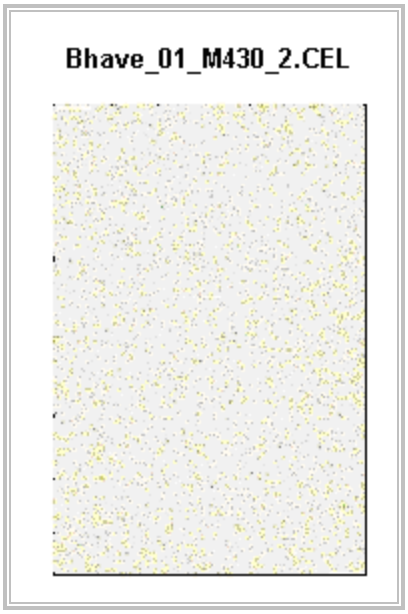
Absolute Deviation of Residuals (MAD) The absolute deviation of the residuals from the median represents deviation of probe level intensities from those predicted, as opposed to the transcript cluster level deviations examined in the preceding Absolute Relative Log Expression graphic. The MAD graphic for exon arrays (below) displays the mean absolute deviation of the residuals from the median for that probe across probes/features within an array and the standard deviation of this value within an array. Consistent values across arrays are ideal. If an array has a higher mean or a significantly larger standard deviation, the quality of this array may be suspect.



The example MAD graph above displays consistent results across samples. The final sample, DBA_3, appears to have a higher mean MAD indicating that the intensities for that particular array deviated to a greater extent from the other arrays. However, small deviation is considered only a minor issue.

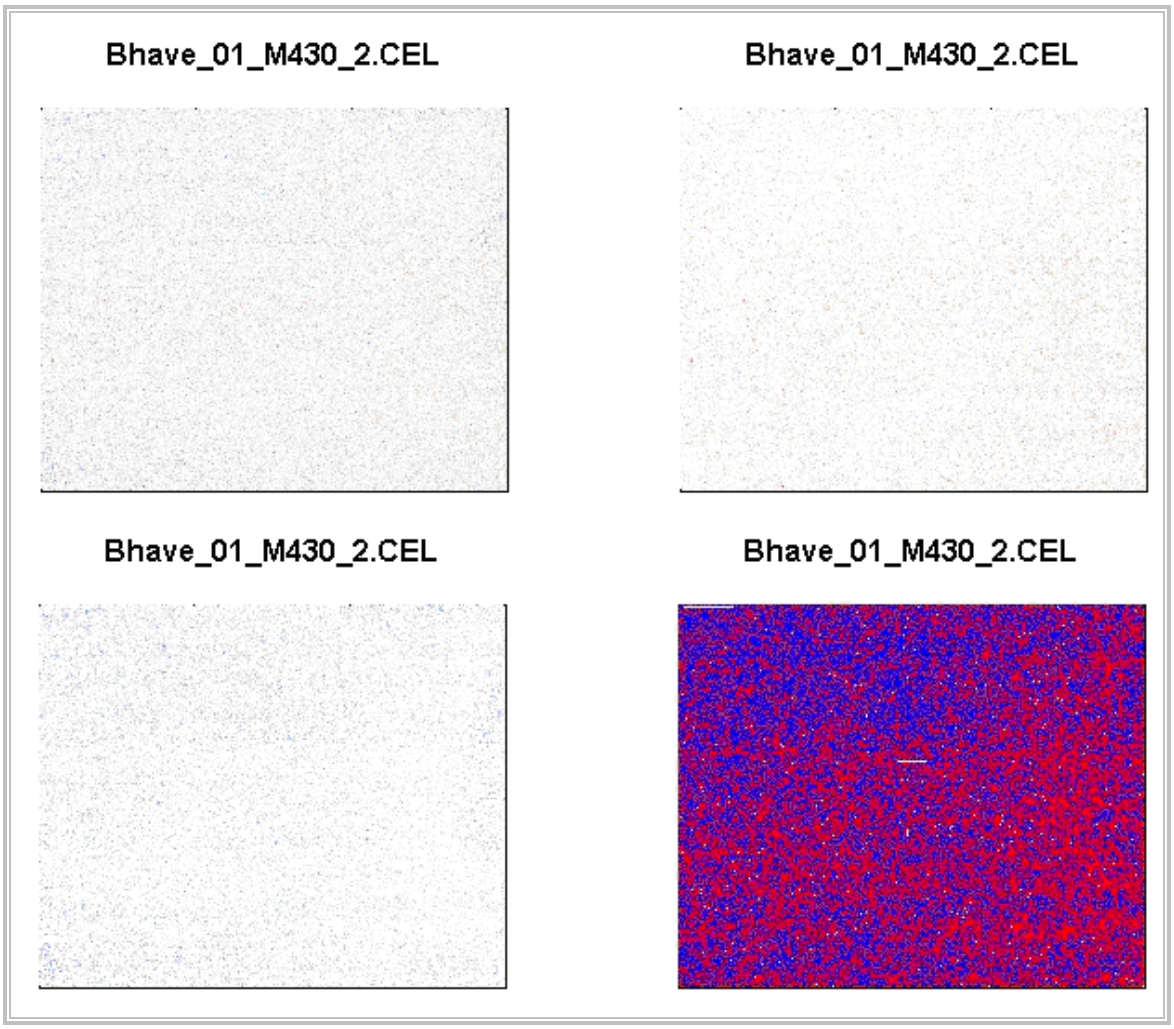
Pseudo Images (Affymetrix)

There are several different pseudo-images that you can inspect for artifacts that are not visible from the raw images. The first image that is displayed shows (from left to right) the spatial distribution of the weights involved in the estimation of the probe-level model outlined previously. In some sense, weights can be considered a "standardized" residual. They range from 0 to 1 where 1 is a small residual and 0 is a large residual relative to the variance between residuals for the probe across all arrays.



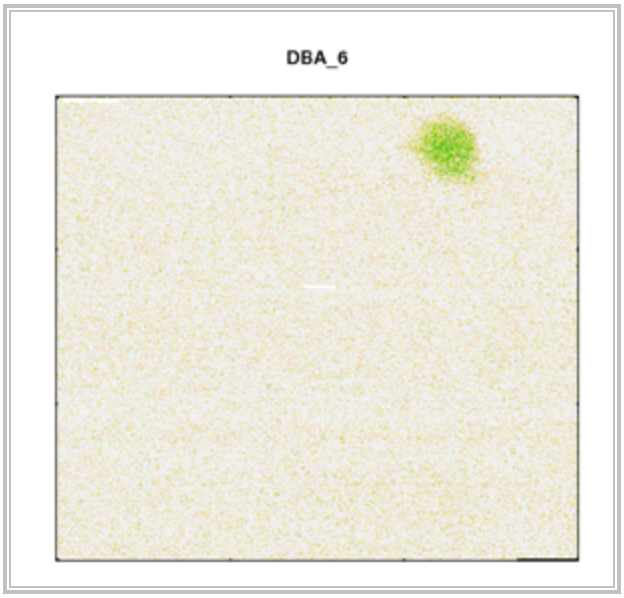
The pseudo-image above looks good because the spots appear to be randomly scattered around the array, rather than being concentrated in any one area. These images are mainly useful for finding spatial artifacts that may be caused by scratches on the array, bubbles that occurred during processing, etc.

You can also look at the raw residuals using a pseudo-image, which represents the εkij value from the default model equation. As opposed to the weights, you want to see values in these images that are close to zero, which indicates that the model is a good fit. There are also several options for looking at the residuals; viewing both the positive and negative residuals on one array or separating them onto two different arrays. You can view a pseudo-image that represents just the sign of the residual, not the magnitude. Following is an example of each choice for the same array.



The image at the top left has both negative (blue) and positive (red) residuals shown. The intensity of each color represents the magnitude of the residual. The image in the top right shows only positive residuals. The image in the bottom left shows only negative residuals. Finally, the image in the bottom right represents the signs of the residuals only. In general, the plots show random distribution of color and intensity, indicating no major artifacts of concern for this array.

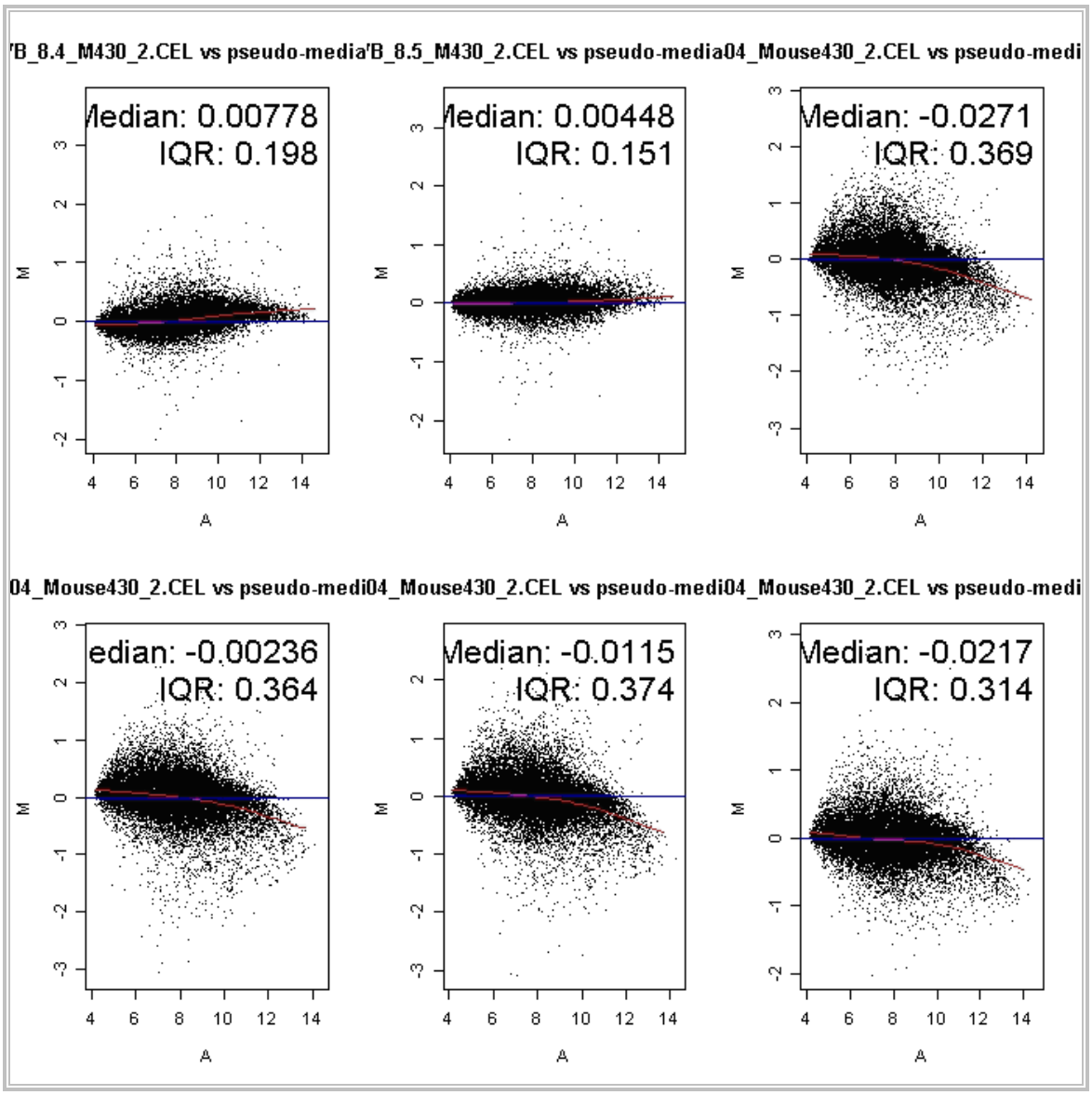
Some pseudo-images show a definitive ‘spot’ on the array.



Although some spots are obvious, Affymetrix requires that ‘artifacts’ must cover over 10% of the array for the array to be considered poor quality. The main reason for this is that the image above displays weights on the probe level. Since probes from the same probeset are scattered about the array randomly, it is assumed that the effect of these ‘bad’ probes will be eliminated when summarized into probeset values during the normalization procedure.

MA Plots

An MA plot is a scatter plot used to compare two arrays. The y-axis is the log-fold change and the x-axis is the average log intensity between the two arrays. The example data uses 20 arrays, so instead of looking at each pair-wise comparison, a "reference" array is used. The reference array in the following graphs is the median intensities across all arrays. The expectation is a random scatter plot, centered about the zero horizontal line. The MA plots that follow are for six arrays.



The blue line in the graph is the zero reference line. The red line is the loess curve, fit to the actual data. The top middle graph shows a "good" MA plot because the points are scattered evenly about the zero reference line and the loess line is close to the zero reference line. The normalized data for the three bottom MA plots shows some biases. Each of the loess curves has a downward slope at the higher average intensities, which indicates that these arrays tend to have lower values than the other arrays in the sample at higher intensities.

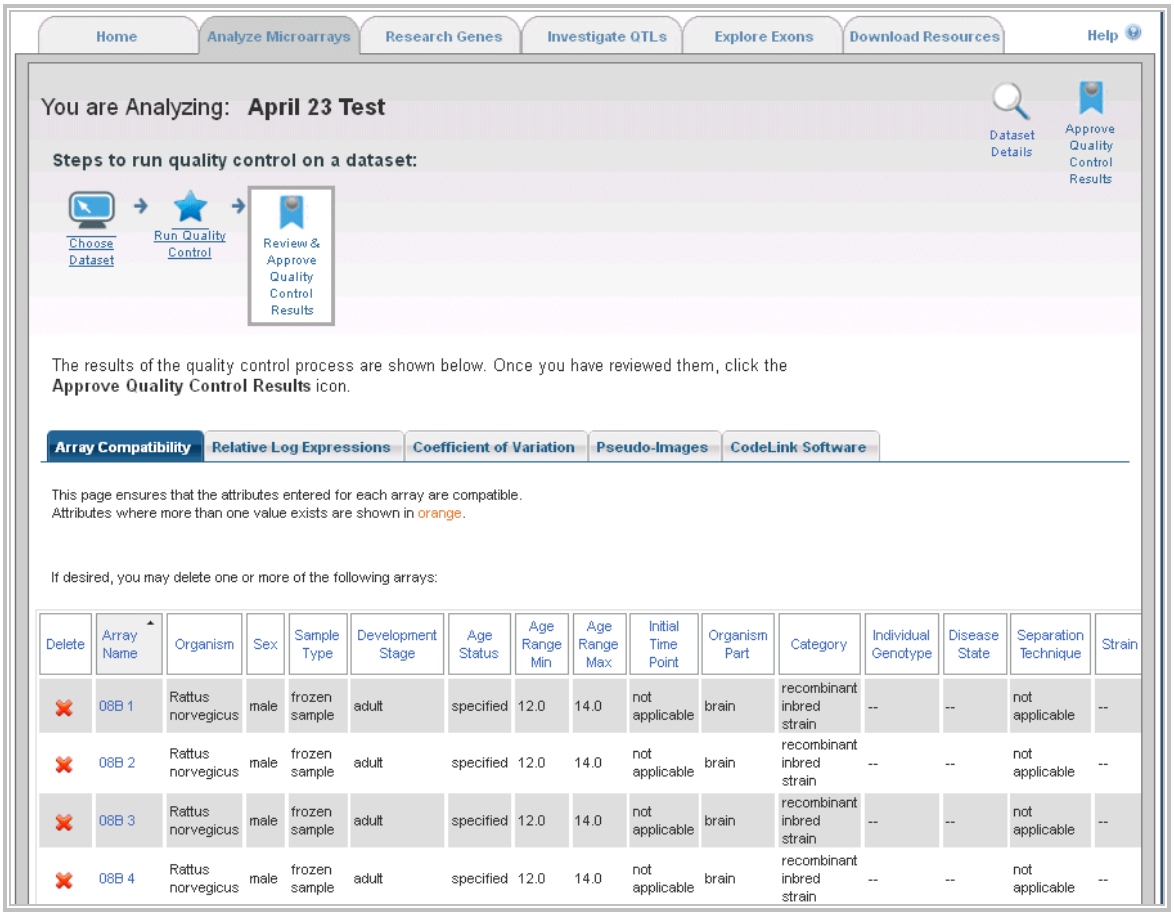
The inter-quartile range (IQR) and median are also reported on the *Quality Control Results* page and each graph. A compact IQR indicates that few genes are different, and there is less variation between arrays. A larger inter-quartile range can indicate that many genes are differentially expressed or that there is more variation between arrays. The IQR for the three bottom graphs is much larger than the IQR for the top center graph. The four graphs with the downward sloping loess curve (bottom three and top right) are arrays that were elevated in the NUSE graph.

Guidelines for Assessing CodeLink Data Quality

Quality control for CodeLink arrays is measured using the distribution of probe intensities, Coefficient of Variation, and a table that displays flags set by the proprietary CodeLink software. After you run quality control on CodeLink datasets, graphs and tables display on individual tabs.

Notes:

- If you choose not to generate images when you run the quality control checks, the *Pseudo Images* tab has no data.
- Click the Download icon to download the images from each tab.



Delete	Array Name	Organism	Sex	Sample Type	Development Stage	Age Status	Age Range Min	Age Range Max	Initial Time Point	Organism Part	Category	Individual Genotype	Disease State	Separation Technique	Strain
✖	08B 1	Rattus norvegicus	male	frozen sample	adult	specified	12.0	14.0	not applicable	brain	recombinant inbred strain	--	--	not applicable	--
✖	08B 2	Rattus norvegicus	male	frozen sample	adult	specified	12.0	14.0	not applicable	brain	recombinant inbred strain	--	--	not applicable	--
✖	08B 3	Rattus norvegicus	male	frozen sample	adult	specified	12.0	14.0	not applicable	brain	recombinant inbred strain	--	--	not applicable	--
✖	08B 4	Rattus norvegicus	male	frozen sample	adult	specified	12.0	14.0	not applicable	brain	recombinant inbred strain	--	--	not applicable	--

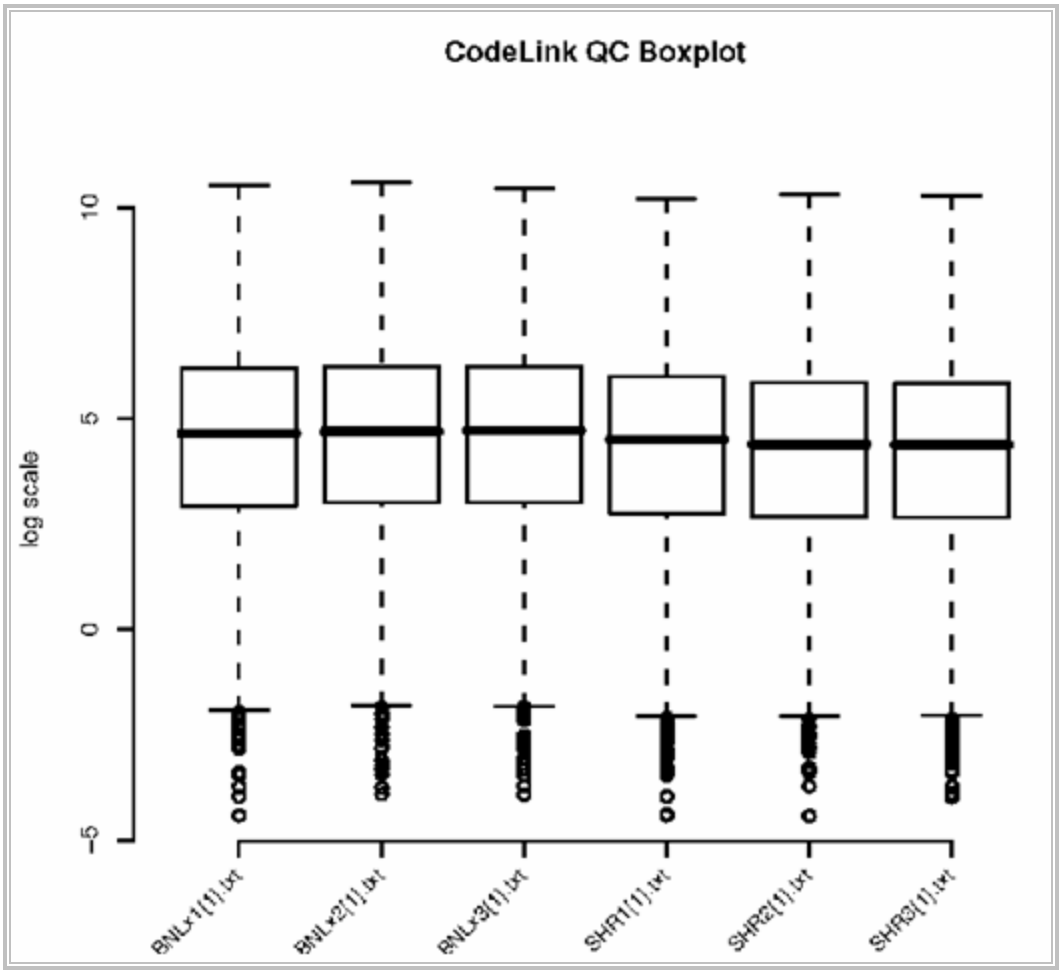
See the following topics for explanations of the data that displays on each tab:

- "Distributions of Probe Intensities" on page 51.
- "Coefficient of Variation" on page 52.
- "Pseudo Images (CodeLink)" on page 52.
- "CodeLink Software" on page 53.

Distributions of Probe Intensities

The distribution of log expression values for each sample are represented as box plots. This quality assessment depends on the assumption that most genes measured are expressed at similar levels across the arrays. The expectation is that the log expression values within a sample should have a similar distribution across samples: e.g., one array does not have a higher median expression value or one array does not show a much wider interquartile range than the others.

In the following image, all six samples show a similar distribution; the boxes are of similar height and the median expression value expressed by the thick horizontal bar within the box are all close to 5. When there is quite a difference between boxes, a quantile normalization or a cyclic loess normalization will force samples to have similar, if not identical, distributions.



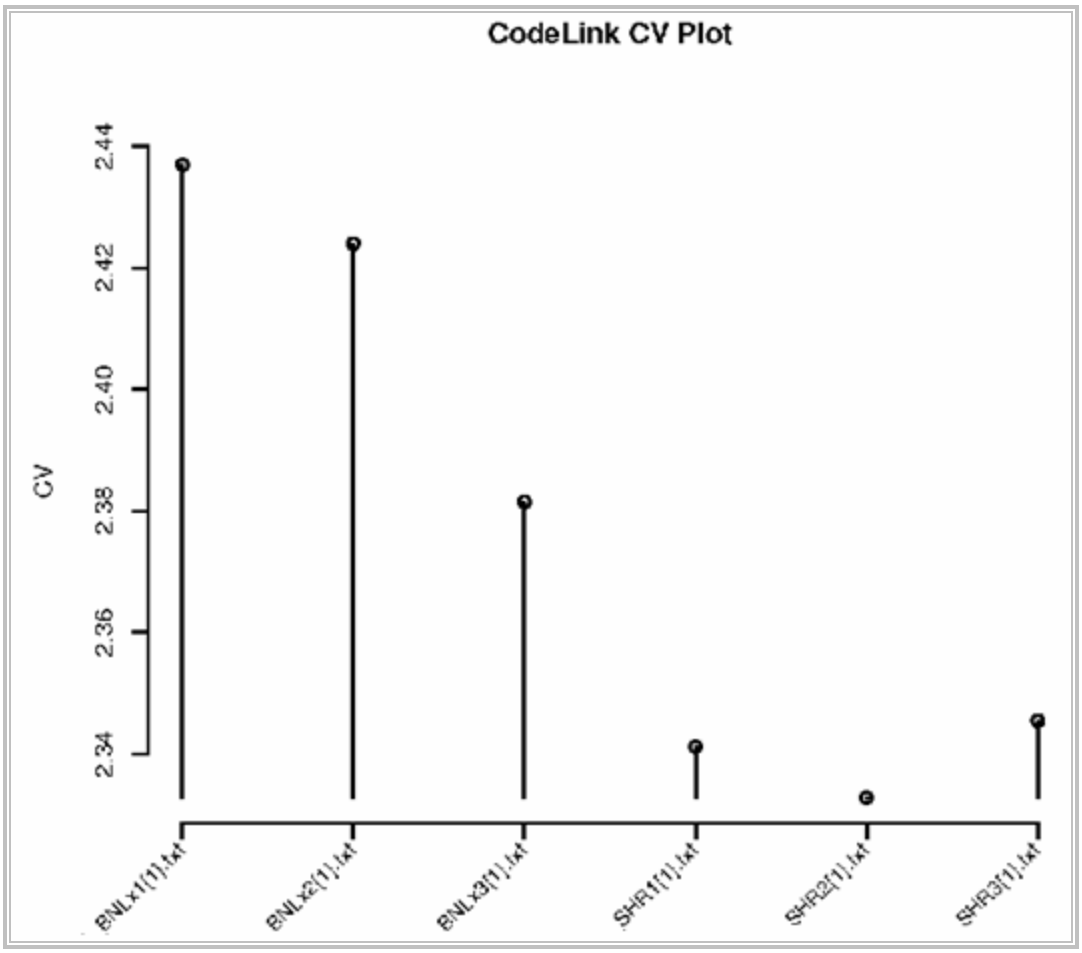
Coefficient of Variation

Coefficient of Variation (CV) is a statistical measure of the variation of data points in a data series around the mean. It is calculated as follows:

CV = standard deviation / mean

In this case, the coefficient of variation represents the ratio of the standard deviation between probe intensities within an array to the mean probe intensity within that array. The CV values are represented on the following graph where a vertical line is dropped down from each CV point to link to its array. You should look for CVs that are similar across arrays, paying particular attention to the range of CVs displayed.

The image that follows shows the CV for six samples from two different strains. Although there is a difference between samples, note that the y-axis runs from 2.34 to 2.44 with only a difference of 0.10 between minimum and maximum values. Differences above 0.5 may require further attention. Also notice that the higher values are all from one strain while the lower values are all from another strain. When values are similar within some biological trait, it is more likely that the differences seen are due to some underlying biological explanation rather than a technical problem.

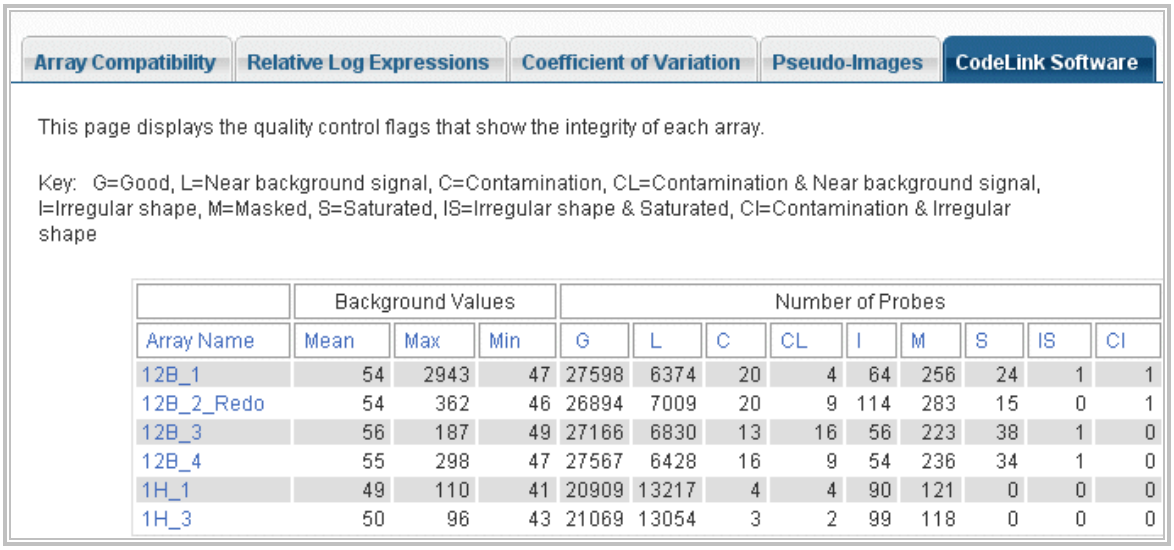


Pseudo Images (CodeLink)

If images are generated during quality control checks, the pseudo-images tab shows a spatial image of the raw spot intensity for each array. When there is a pseudo-image generated, you should look for any major artifacts that would indicate an area of the array that has been compromised during the hybridization process. For examples of good and bad pseudo images, see "Pseudo Images (Affymetrix)" on page 46.

CodeLink Software

CodeLink proprietary software produces quality control flags that show the integrity of each probe on each array. These flags may serve as an indication of inferior arrays. The table lists summary statistics for each array, as well as the number of spots labeled with CodeLink Calls values. For example, arrays having high number of CodeLink Calls other than G (Good), compared to the other arrays within the dataset, can be considered for elimination. See "CodeLink Gene Filtering Procedures" for details about the Number of Probes key (G=Good, L=Near bg signal, C=Contamination, CL=Contamination & Near bg signal, I=Irregular shape, M=Masked, S=Saturated, IS=Irregular shape & Saturated, CI=Contamination & Irregular shape).

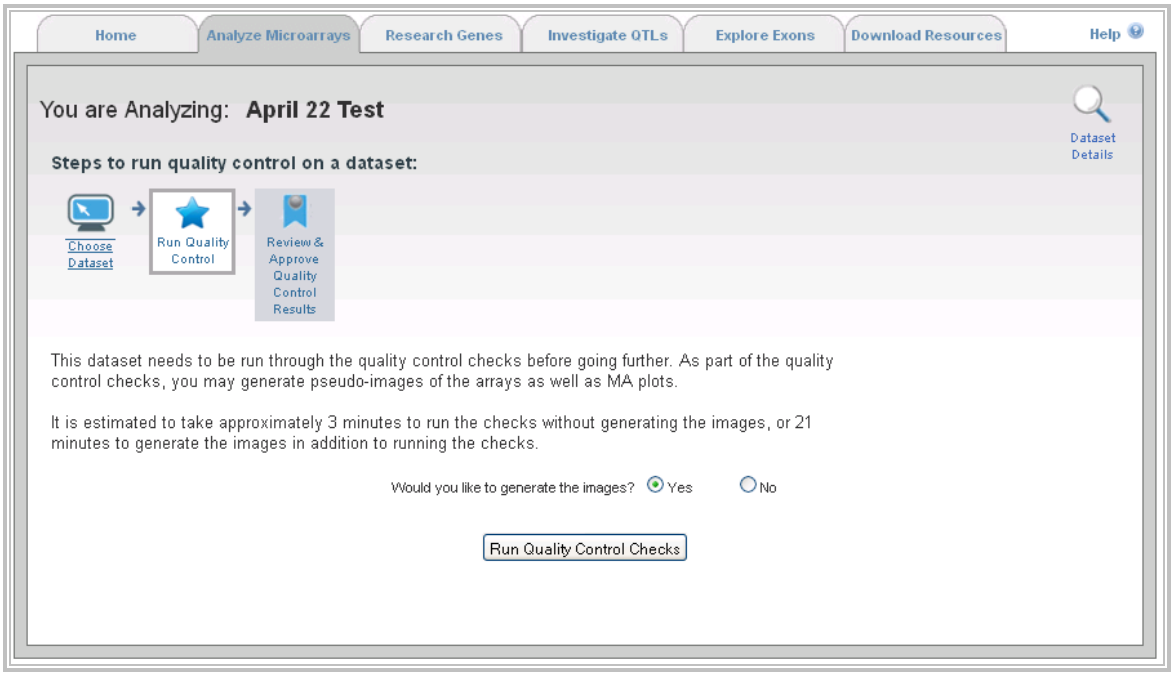


	Background Values			Number of Probes							
Array Name	Mean	Max	Min	G	L	C	CL	I	M	S	IS	CI
12B_1	54	2943	47	27598	6374	20	4	64	256	24	1	1
12B_2_Redo	54	362	46	26894	7009	20	9	114	283	15	0	1
12B_3	56	187	49	27166	6830	13	16	56	223	38	1	0
12B_4	55	298	47	27567	6428	16	9	54	236	34	1	0
1H_1	49	110	41	20909	13217	4	4	90	121	0	0	0
1H_3	50	96	43	21069	13054	3	2	99	118	0	0	0

Running a Quality Control Check

After you create a dataset, you must run a quality control check on it. Arrays that are identified as questionable at any of the steps should be considered for deletion. However, some of the small imperfections and minor concerns can be alleviated by an appropriate normalization method. See "Preparing Datasets".

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets. Datasets that require quality control checks show *Run* in the **QC Complete** column.
3. Click the dataset on which you want to perform quality control. The *Quality Control* page displays.
4. Choose whether you want to generate pseudo images from the QC run. If you do not generate images, the **Pseudo Images** and **MA Plot** (Affymetrix) tabs in the quality control results page do not display results.
5. Click **Run Quality Control Checks**. A confirmation message displays. Click **Close**.



The quality control checks take time, especially when you generate images. When the checks are complete, an email is sent to the address you provided in the *Registration* page, and you can view the results. If the quality control process encounters errors, you must revise your array selection and re-run the quality control process. See "Viewing and Approving Quality Control Results" for instructions on deleting chips.

Viewing and Approving Quality Control Results

After you run the quality control checks on your datasets, you must view and approve the results.

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.

   Each dataset has one of these designations in the **QC Complete** column, to indicate the stage of the quality control process:

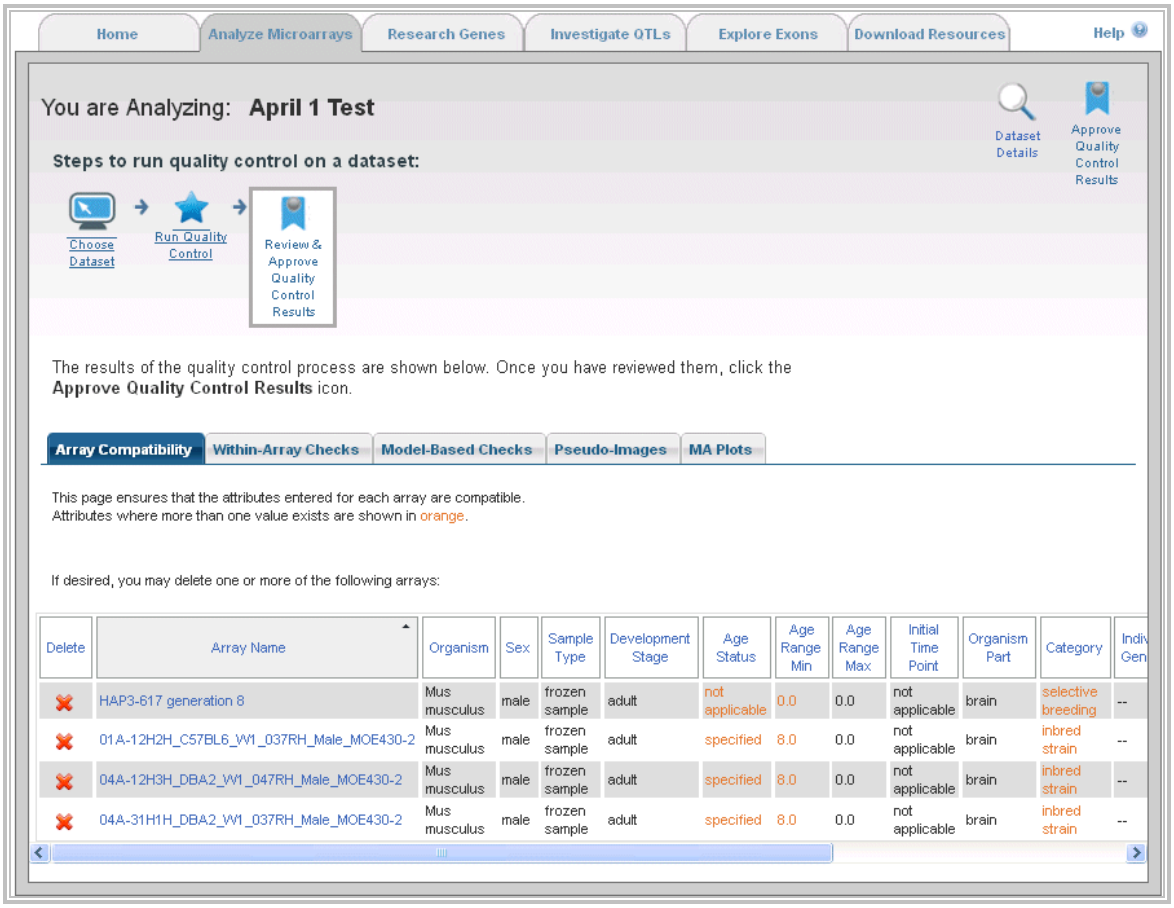
   - *Run:* The quality control checks need to be run.
   - *In Progress:* The quality control checks are running.
   - *Review Results:* The results of the quality control checks are available for review.
   - *Checkmark:* The results of the quality control checks are approved.
3. Click a dataset with "Review Results" in the **QC Complete** column to review the results of the quality control checks.
4. Review the information displayed on each tab. If appropriate, click the **Delete** icon to delete an array from the dataset. If you delete an array, you must re-run the quality control process.
5. Click the **Approve Quality Control Results** icon to approve the results.

Note: To see the approved Quality Control Result for a dataset, click the **Analyze Microarrays** tab, click **Analyze microarray data** in the *What would you like to do* section, then click the **View Details** icon in the **QC Results** column.

Quality Control Results

The quality control results page displays tabs to separate the types of quality control checks. The **Array Compatibility** tab displays the results of the Array Attribution Comparison. The remaining tabs display the results of the Array Integrity Checks.The tabs that display depend on whether the dataset contains CodeLink or Affymetrix data. See "Quality Control Checks Overview" for details.

Each tab after the **Array Compatibility** tab displays a graph of the results. If you did not choose to display images when you ran the quality control check, the **Pseudo Images** tab and the **MA Plot** tab (Affymetrix only) do not contain images. See "Guidelines for Assessing Affymetrix Data Quality" and "Guidelines for Assessing CodeLink Data Quality" on page 50 for an explanation of the data that displays on each tab.



Note: Click the Download icon to download the images from each tab.

Preparing Datasets

After you finalize a dataset, run the quality control checks on the dataset, and approve the quality control results, your dataset is ready to be grouped and normalized, as indicated by the word "Run" in the **Arrays Grouped and Normalized** column on the *View Datasets* page.

Grouping

Data grouping allows you to group arrays and normalize data across groups. Data is then in a state where statistical analysis can be performed. You can create multiple groupings and normalize each grouping multiple times. After you create and normalize a grouping of arrays, a checkmark displays in the Arrays Grouped and Normalized column.

Groups

The term "group" is used to indicate an analysis group. For example, if you analyze mouse data for differential expression between males and females, group 1 can be all the samples from female mice, and group 2 can be all the samples from male mice. For inbred strains, replicate samples within a group can be considered biological replicates, because even though the samples are two different animals, it is assumed that the gene expression is similar (if not exactly the same) between the two samples.

You can choose groups by:

- Array attribute.
- Previous saved group combinations.
- User-created categories.

Normalization

The purpose of normalization and background correction is to remove systematic noise and reduce technical variation. To normalize in the context of DNA microarrays means to standardize your data to be able to differentiate between real (biological) variations in gene expression levels and variations due to the measurement process. Normalizing also scales your data so that you can compare relative gene expression levels. In general, the normalization process is subdivided into four sequential steps:

1. Background correction.
2. Data normalization.
3. Adjustment for non-specific binding (Affymetrix arrays only).
4. Data summary methods (Affymetrix arrays only).

The options for normalizing data are based on array platform:

Affymetrix 3' Arrays	Affymetrix Exon Arrays	CodeLink Arrays
○ MAS5.0 ○ dChip ○ RMA (recommended) ○ VSN ○ GCRMA	○ RMA ○ PLIER	○ None ○ Loess (recommended) ○ VSN ○ LIMMA

Many normalization methods have been developed over the years since microarrays first hit the market. These different normalization methods have the potential of yielding very different results in candidate gene searches. Some researchers choose to run their analysis using a couple different normalization methods and then choose the candidate genes that are identified regardless of normalization method. For Affymetrix, RMA

and the closely related gcRMA are the most common normalization methods published. For CodeLink, the paper cited in the following *References* section gives an in-depth comparison of normalization methods and concludes cyclic Loess to be the most accurate.

References

Wu W, Dave N, Tseng GC, Richards T, Xing EP, and Kaminski N (2005). Comparison of normalization methods for CodeLink Bioarray data. BMC Bioinformatics 6:309.

Normalization Methods

MAS 5.0

Micro Array Suite Version 5.0 (MAS5) is implemented in both the MAS 5.0 software package from Affymetrix and in the *Affy* package in R. The PhenoGen website application uses the *mas5* function in R. Signal calculation using MAS 5.0 consists of five main steps.

The first step adjusts the raw intensities for a global background by organizing the array into zones. Within each zone, the lowest 2% of intensities are used as an estimate background for that zone. The transition between zones is smoothed by taking a weighted estimate of background for each point where the weights are based on distances from zone centers. The second step calculates an "ideal" mismatch (IM) intensity to adjust the perfect match intensity with the goal of eliminating background cross-hybridization and stray signal. The IM value is used instead of the mismatch (MM) value to ensure that the resulting signal (PM-IM) is positive. The third step transforms the intensity values with a log base 2 transformation. The fourth step combines probe values within a probe set using the one-step Tukey's biweight algorithm, which "weights" the data to reduce the influence of outliers. The fifth and final step scales all probe sets after conversion back to the original intensity scale, i.e., not log base 2 transformed, to a target probe set intensity (500 by default). After the MAS 5.0 procedure is complete, all intensity values are transformed using a log base 2 in preparation for statistical analyses.

References

1. Affymetrix (2001). Statistical algorithms reference guide. Technical report Affymetrix.
2. Hubbell E, Liu W-M, Mei R (2002). Robust estimators of expression analysis. Bioinformatics 18(12):1585-1592.
3. Liu W-M, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho M-H, Baid J, Smeekens SP (2002). Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics 18(12):1593-1599.

PLIER (Probe Logarithmic Intensity Error)

Probe Logarithmic Intensity Error Estimation (PLIER). The PLIER algorithm is a model-based method that includes "experimentally observed patterns for feature behavior and handling error appropriately at low and high abundance". The probe, or "feature", intensities are first pre-processed by quantile normalization to scale expression values across arrays. Intensities are then adjusted for background noise by subtracting the median intensity of control probes with similar GC content. All feature intensities for a probe set are used to calculate a feature response; "a measure of how much the relative intensity of a feature is due to the feature itself, as opposed to the common target of a probe set". After taking into account the difference in features within a probe set, intensity values across features are combined using a weighting scheme that down-weights probes with inconsistent behavior. Finally, a mixed error model is used to account for the differences in appropriate error models depending on the abundance of the transcript.

References

1. Guide to Probe Logarithmic Intensity Error (PLIER) Estimation, Affymetrix Tech Note, 2005.

dChip (Perfect Match Probes Only)

DNA-chip Analyzer (dChip) is a software package that implements the model proposed by Cheng Li and Wing Hung Wong. This model is also referred to as the Li Wong method and the resulting values are considered model-based expression indexes (MBEI). The PM-only model fits the following model to each gene:

$$PM_{ij} = v_j + \theta_i \phi_j + \varepsilon_{ij}$$

Where:

PM_{ij} is the intensity for perfect match probe of probe pair *j* on the *i*th array.

v_j is the baseline response of the *j*th probe due to nonspecific binding.

θ_i is the expression index of the gene in the ith array.

ϕ_j is the sensitivity of the PM probe of the probe pair *j*.

ε_{ij} is the random error.

Parameter estimates are determined through iteration. The R code used for normalization on the PhenoGen website is not identical to the code used in Li and Wong's stand-alone program, *dChip*. Therefore, normalization on the website might differ slightly from normalization derived from the *dChip* program. After the dChip method is completed, all intensities are transformed using a log base 2 in preparation for statistical analyses.

References

1. Li C, Wong WH (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98(1):31-6. (uses both PM and MM probes)
2. Li C, Wong WH (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues, and standard error application. Genome Biology 2(8):1-11. (uses only PM probes)

RMA

Robust Multi-Chip Average/Robust Multi-Array Analysis (RMA). RMA pre-processing consists of three main steps: background correction, normalization, and summarization of probe level intensities in probe sets. RMA uses a background correction method to account for optical noise and non-specific binding using only the perfect-match probes. The background-corrected probe intensities are then transformed using log base 2 and normalized using quantile normalization. Finally, the probe intensities are combined using a median polish to get one intensity for each probe set or transcript cluster. The log base 2 transformation of intensity values occurs within the RMA procedure, so a separate transformation is not needed.

For the Affymetrix Exon Arrays, determine the probesets to include, based on confidence in annotation (core, extended, and full), by summarizing on either the exon level or gene level. For more details, see the Affymetrix GeneChip ® Exon Array whitepaper, *Exon Probeset Annotations and Transcript Cluster Groupings (2005)*.

References

1. Bolstad BM, Irizarry R A, Astrand M, and Speed TP (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193.
2. Irizarry RA, Benjamin M, Bolstad FC, Cope LM, Hobbs B, Speed TP (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 31(4):e15.
3. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics 4(2): 249-264.

VSN

Variance stabilization (VSN). The goal of variance stabilization is to transform the data in such a way as to eliminate the dependence of the variance on the mean. The VSN accounts for both background correction and normalization by scaling and shifting the original intensity and then using the inverse of the hyperbolic sine as a variance stabilizing transformation.

For Affymetrix data, this normalization method is implemented in the Affy *vsnrma()* procedure. The VSN transformation is done at the probe level on the PM probes only, and then the probe level intensities are combined into a probe set intensity using a median polish. Affymetrix data is transformed to log base 2 within the VSN procedure.

For CodeLink data, this normalization method is implemented by choosing either the 'vsn' or the 'limma' options on the website. If the 'vsn' option is chosen, it is implemented using the *vsn2()* procedure in R and the values are on the natural log scale. If the 'limma' option is chosen, it is implemented using the *normalizeBetweenArrays()* procedure, and values are on the log base 2 scale. In either case, data values are background corrected before VSN is implemented.

Finally, for both Affymetrix and CodeLink data, a quantile of 0.5 is used for the least-trimmed sum of squared (LTS) regression for the estimation of parameters. This is the most robust value (i.e., a value of 1 offers no protection to outliers and 0.5 offers the most).

References

1. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18(S1):S96-S104.
2. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2003). Parameter estimation for the calibration and variance stabilization of microarray data. Statistical Applications in Genetics and Molecular Biology 2(1) Article 3. http://www.bepress.com/sagmb/vol2/iss1/art3

GCRMA

G/C Robust Multi-Array Average (GCRMA). The GCRMA method is related to the RMA method described previously, in that GCRMA also uses quantile normalization on the log base 2 probe values and a median polish to summarize probes into a probe set. However, the difference lies in the background adjustment. The RMA method does not account for probe affinity when calculating background. For GCRMA, the estimate for nonspecific binding is related to the base composition of the nucleic acid molecules, i.e., the proportion of G and C bases present in the probe sequence.

References

1. Wu Z, Irizarry RA, Gentleman R, Martinez Murllio F, Spencer F (2004). A model based background adjustment for oligonucleotide expression arrays. Johns Hopkins University, Dept. of Biostatistics Working Papers, Paper 1. http://www.bepress.com/jhubiostat/paper1

LOESS

Locally Weighted Scatterplot Smoothing (LOWESS/LOESS). When applied to normalization of CodeLink arrays, this method is also referred to as cyclic loess. This method is an iterative approach that is based on the MA plot. For each distinct pair of arrays, the data is plotted using an MA plot (difference in log base 2 values versus the average of log base 2 values). A loess curve using a one-degree polynomial is fit to each graph. This loess curve is used to estimate an adjustment for each value. The average adjustment for all pairwise comparisons for a particular probe is used to obtain starting values for that probe and array in the next iteration. This method is implemented using the *normalize.loess* function in R. Initially, expression values from CodeLink are adjusted for background, i.e., spot mean-background median. Because this can result in

negative values, expression values less than 1 are re-coded to 1. Expression values are transformed using a log base 2 before LOESS normalization. You can fit the loess curve using a weighted least-squares approach (family.loess=gaussian) or a re-descending M estimator with the Tukey's biweight function (family.loess="-symmetric").

References

1. Bolstad BM, Irizarry R A, Astrand M, Speed TP (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193.

LIMMA

Linear Models for MicroArrays (LIMMA). LIMMA refers to the package that is used in R. Two normalization methods are available under this method: **scale** and **quantile**.

- *Scale* – Data is background corrected, values less than 1 are re-coded to 1, and all values are log base 2 transformed. Values are then scaled so that the median absolute deviations (MADs) are the same across arrays.
- *Quantile* – Data is background corrected, values less than 1 are re-coded to 1, and all values are log base 2 transformed. Values are then adjusted so that each array has the exact same distribution of intensities. This gives the same results as the **normalize.quantiles** function in R.

References

1. Bolstad BM, Irizarry R A, Astrand M, Speed TP (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193.
2. Smyth GK, Speed TP (2003). Normalization of cDNA microarray data. Methods 31(4): 265-273.

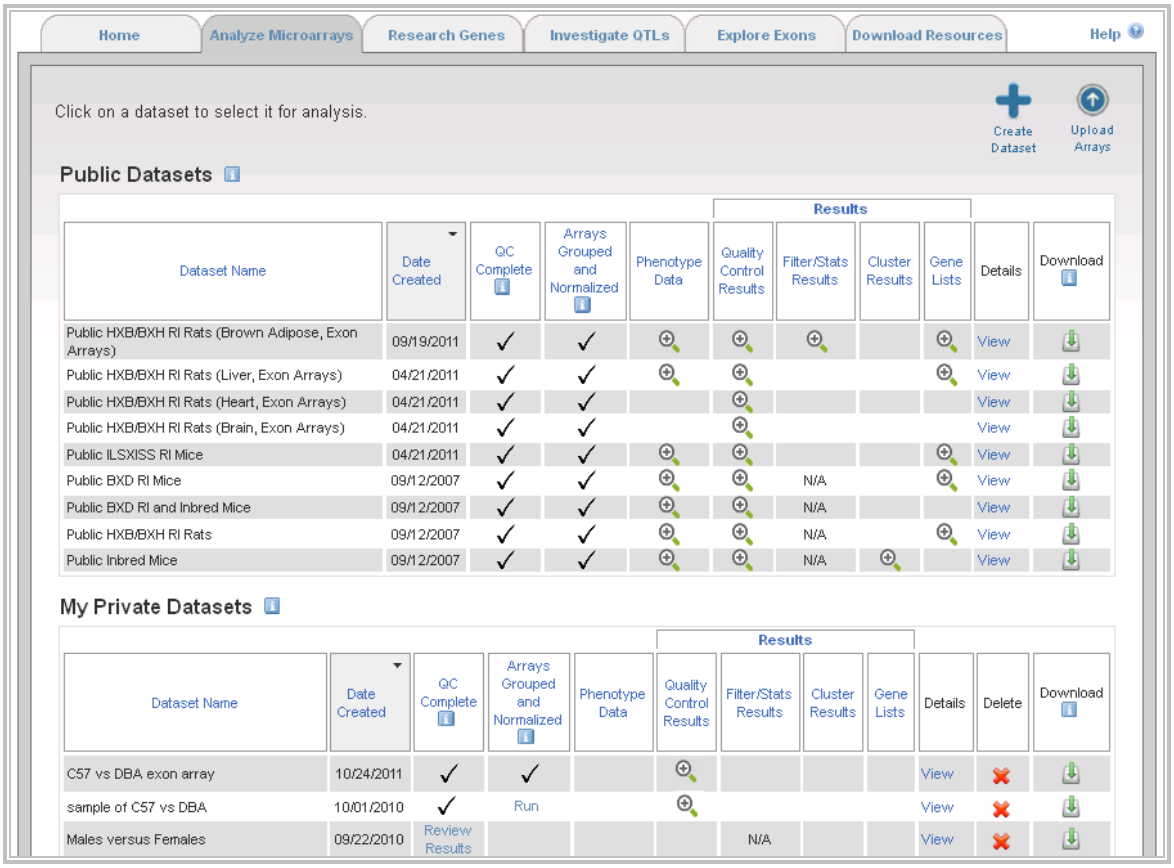
Eliminating Probes with Poor Sequence Integrity

Probe masks were created for the Affymetrix Mouse 430 version 2 array, the Affymetrix Mouse Exon array, and the Affymetrix Rat Exon array to eliminate probes whose sequence did not match to the reference genome (NCBI m37 build for the mouse arrays and RGSC version 3.2 for the rat array), matched the genome in multiple places, or harbored a SNP. For the mouse arrays, SNPs were derived from the 19 strains in the public Inbred Mice dataset where genotype data is available at the Imputed Genotype Resource from the Jackson Laboratory; http://cgd.jax.org/datasets/popgen/imputed.shtml (129P3/J is not available). For the rat array, SNPs were derived from comparing the full genome sequence of the Brown Norway (BN) Inbred strain to the Spontaneously Hypertensive Rat (SHR) strain that was recently sequenced (Atanur et al 2010) using next-generation sequencing. We also included SNPs identified from our sequencing of the DNA (Phred quality score > 150) of the parental strains of the HXB/BXH panel (SHR/OlaPrin and BN-Lx/CubPrin).

Entire probe sets were eliminated if less than four probes remained for the 3' array or less than three probes remained for the exon array. For the Affymetrix Mouse 430 version 2 array, 68,002 probes were eliminated because they did not align uniquely to the genome, and 39,430 additional probes were eliminated because they targeted a genomic region with a known SNP. Of the original 45,101 probe sets on the array, 41,485 remain after masking. For the Affymetrix Mouse Exon array, 329,422 probes were eliminated because they did not align uniquely to the genome, and 913,592 additional probes were eliminated because they targeted a genomic region with a known SNP. Of the original 1,180,331 probe sets from the full annotation on the array, 900,079 remain after masking. For the Affymetrix Rat Exon array, 472,072 probes were eliminated because they did not align uniquely to the genome, and 132,529 additional probes were eliminated because they targeted a genomic region with a known SNP. Of the original 887,561 probe sets on the array from the full annotation, 721,150 remain after masking.

Grouping and Normalizing Datasets

Your dataset is ready to be grouped and normalized, when the word *Run* displays in the **Arrays Grouped and Normalized** column on the *View Datasets* page.



1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.
3. Click the dataset you want to group.
4. Select the parameters of the group:
   - **Create groups automatically:** Choose a criterion from the **Create groups based on the following** drop-down list. The list contains attributes whose values differ amongst the arrays in the dataset, and the samples are sorted into groups automatically. If your experiment contains arrays from replicate lines, *Replicate Dataset* displays in the list.

     OR

   - **Create a group manually:** Manually add samples to the **Group** or **Exclude** columns.

   Notes:

   - Click the **Create Additional Group** link, at the top right of the table, to create a new group.
   - Click **Name this Group** in the column headings to enter a descriptive name.
5. Enter a name for the grouping, and click **Next**.
6. Select the grouping you want to use for this normalized version from the list.
7. Select a Normalization Method from the drop-down list. See "Normalization" for details.

Normalizing Affymetrix Arrays

If you have an experiment with Affymetrix arrays, choose one of the following options and proceed to step 8:

- **MAS5** (method implemented in Affymetrix GeneChip 5 software).
- **dChip** (DNA chip analyzer).
- **rma** (robust multi-array average method).
- **vsn** (variance stabilization normalization method).
- **gcrma** (G/C robust multi-array average).
- **plier** (Exon arrays only)

Normalizing CodeLink Arrays

If you have an experiment with CodeLink arrays, choose one of the following options and proceed to step 8:

- **None**.
- **Loess** (Locally Weighted Scatterplot Smoothing) - Select **gaussian** or **symmetric** from the drop-down list that displays.
- **vsn** (variance stabilization normalization method) - Select **Lts.quantile** and **Number of iterations** from the drop-down lists that display.
- **LIMMA** (Linear Models for MicroArrays) - Select the **Limma method** from the drop-down list that displays.

8. Enter a **version name** for this normalized version of the dataset.
9. Click **Next**. The normalization process takes time. When it is complete, you receive an email.

! **Important!** You can group and normalize the same dataset any number of ways. Each is saved .

Analyzing Datasets

A "dataset" in the PhenoGen website is created using a collection of arrays from one or more lab experiments. You can create datasets using arrays in the PhenoGen database that are public or that you have been granted access to use. There are also four pre-compiled "Public" datasets. These can be used by all registered users for correlating phenotype data with gene expression data in inbred and recombinant inbred mice and rat strains.

After you group and normalize a dataset, you can analyze it. Data analysis consists of:

1. Filtering out noise by applying optional filters that allow you to eliminate certain probe sets from further analysis.
2. Identifying statistically significant genes using statistical analysis tools.
3. Saving the resultant set of probes or probesets as a gene list.

The above steps can be further broken down into the following process:

1. Select a single normalized dataset version.
2. Choose the type of analysis to perform:

Differential Expression

- Proceed to Step 3, Filter genes.

Correlation Analysis

a. Create a behavioral phenotype data file off line. See "Uploading Phenotype Data" for the required format.

b. Upload the phenotype data file.

c. Proceed to Step 3, Filter genes.

Clustering

- Proceed to Step 3, Filter genes.

3. Filter genes.
4. Perform statistical analysis or clustering.
5. Save gene lists or cluster results.

Filtering

A typical microarray consists of thousands of probe sets (10,000 - 1,00,000), and the introduction of meaningless noise is inevitable. Removing this noise increases the chances of finding significant genes. The PhenoGen website provides various gene filtering methods for filtering out genes prior to running statistical analysis. For differential expression analysis and correlation analysis, the filters are based on the types of arrays used in the dataset. See:

- "Affymetrix Gene Filtering Procedures" on page 63
- "CodeLink Gene Filtering Procedures" on page 65

For clustering analysis, the filters are also based on the types of arrays in the dataset, but additional filters based on gene expression values are also available. See:

- "Clustering Filtering Procedures" on page 67

Note: The PhenoGen website does not provide custom array filtering.

Affymetrix Gene Filtering Procedures

Affy Control Gene Filter

Use the *Affy Control Gene Filter* to remove internal control probes from the analysis. Affymetrix technology uses house-keeping genes' intensity values as a means for quality control. A pre-determined concentration of these control genes was spiked into the cRNA target mixture prior to application onto the microarrays. The measured intensity values are used for internal quality control. Since these control genes are typically from different species, they have little importance to the analysis.

MAS5 Absolute Call Filter

Use the *MAS5 Absolute Call Filter* to either keep or remove probes based on their present or absent call. The MAS5 algorithm uses probe-pair intensities to generate a detection p-value and assign a Present, Marginal, or Absent call. Each probe pair in a probe set has a potential vote in determining whether the measured transcript is detected (Present) or not detected (Absent). The vote is described by a value called the discrimination score [R]. The score is calculated for each probe pair and is compared to a predefined threshold Tau. Probe pairs with scores higher than Tau vote for the presence of the transcript. Probe pairs with scores lower than Tau vote for the absence of the transcript. The voting result is summarized as a p-value. The greater the number of discrimination scores calculated for a given probe set that are above Tau, the smaller the p-value and the more likely the given transcript is truly present in the sample. The p-value associated with this test reflects the confidence of the detection call.

The detection p-value cut-offs, alpha 1 (α1) and alpha 2 (α2), provide boundaries for defining Present, Marginal, or Absent calls. At the default settings determined for probe sets with 16 - 20 probe pairs (defaults α1 = 0.04 and α2 = 0.06), any p-value that falls below α1 is assigned a Present call, and above α2 is assigned an Absent call. Marginal calls are given to probe sets which have p-values between α1 and α2.

DABG Absolute Call Filter

Analogous to the MAS5.0 present/absent calls from the Affymetrix 3' Array, each probe set on the Affymetrix Exon array is given a p-value associated with the hypothesis that the intensity values can be distinguished from background noise. This p-value is referred to as the detection above background (DABG). It is generated by comparing each probe in the probe set to a set of background probes with similar GC content. The probe-level p-values are combined into a DABG p-value on the probe set level. For this filter, probe set with a p-value less than 0.0001 is considered "present" and a probe set with a p-value greater than or equal to 0.0001 is considered "absent".

Heritability Filter

Use the *Heritability Filter* to limit the probe sets analyzed to those with a high genetic heritability. This filter eliminates probe sets for transcripts whose environmental influence on expression is high compared to the strict genetic influence. For analyses on the Affymetrix Mouse 430 version 2 array, the broad sense heritability has been calculated on the public inbred mouse panel (20 strains) normalized using RMA with poor quality probes eliminated prior to normalization and the public BXD recombinant inbred mouse panel (32 strains) normalized using RMA with poor quality probe eliminated prior to normalization.

For analyses on the Affymetrix Mouse Exon array, the broad sense heritability has been calculated on the core transcript clusters from the public LXS brain recombinant inbred mouse panel normalized using RMA with poor quality probes eliminated prior to normalization. For analyses on the Affymetrix Rat Exon array, the user must choose the tissue of interest (brain, heart, liver, or brown adipose tissue). The broad sense heritability has been calculated separately for each tissue on the core transcript clusters from the public HXB/BXH recombinant inbred rat panel normalized using RMA with poor quality probes eliminated prior to normalization.

The broad sense heritability is calculated for each probe set/transcript clusters separately using an ANOVA model. Because the public data sets are based on the Affymetrix Mouse 430 version 2 array, the Affymetrix Mouse Exon arrary, and the Affymetrix Rat Exon array, the Heritability filter is only available for data sets on these chips. Use either panel's heritability values for this filter and specify a minimum heritability threshold for inclusion. All filtering is done on the probe set or transcript cluster level.

eQTL/bQTL Filter

Another way to prioritize genes for analysis is to limit those considered to be genes whose transcription levels are controlled from the same genetic region that controls the phenotype/behavior of interest (e.g., Tabakoff et al. 2009). We have identified expression quantitative trait loci (eQTL) for probe sets from the BXD recombinant inbred panel on the Affymetrix Mouse 430 version 2 array, for core transcript clusters from brain tissue of the LXS recombinant inbred mouse panel on the Affymetrix Mouse Exon Array, and for core transcript cluster from brain, heart, liver, or brown adipose tissue of the HXB/BXH recombinant inbred rat panel on the Affymetrix Rat Exon Array.

When data sets are created based on any of these three array technologies, the respective eQTL data sets are used. When using the HXB/BXH recombinant inbred rat panel, the user must choose the correct tissue. The user also chooses a significance threshold for eQTL and the appropriate bQTL to compare to. Probe sets/transcript clusters are retained if their eQTL is significant and the location of the marker (SNP) with the maximum association with transcript expression is within the chosen bQTL limits. All filtering is done on the probe set or transcript cluster level.

Gene List Filter

Filtering by Gene List allows you to select a gene list that has been previously created and either keep or remove all the genes within that gene list.

CodeLink Gene Filtering Procedures

CodeLink Control Gene Filter

There are three basic types of CodeLink probes that are used in the Control Gene Filter:

- **(D) Discovery** - The probes corresponding to the genes of interest for a particular species.
- **(P) Positive controls** - The bacterial probes corresponding to bacterial transcripts that are spiked in at the total RNA level and are used to evaluate the sensitivity and dynamic range of the platform.
- **(N) Negative controls** - The bacterial probes used for evaluating the degree of non-specific assay background and negative control threshold.

When this filter is complete, both P and N are removed.

CodeLink Call Filter

Use the *CodeLink Call Filter* to remove internal control probes from the analysis. The CodeLink Call filter is based on the quality control flag results measured from the imaging software; *CodeLink Expression v4.1 algorithm*. The flags are:

Flag	Description
G	The spot has passed all quality control measures and is defined as good.
M	The spot is identified in the MSR (Manufacturing Spot Removed) File and no intensity data is provided. The probe was masked after printing because it represented a suboptimal probe. Data from these spots is disregarded.
C	The spot has a high level of background contamination. Its background is above the global background population.
I	The spot has an irregular shape.
L	The spot has a near background signal.
S	The spot has a high number of saturated pixels, typically above 60,000 units.

A spot may receive more than one quality control flag. For example, it may be labeled CI if it has both background contamination and an irregular shape.

For the filter, you can specify the number of samples with either a G (Good) flag or an L (Low) flag needed, to keep or remove the probe from further analysis. Samples that have any of the other flags are not counted towards the number of samples needed for retention or elimination.

GeneSpring Call Filter

The *GeneSpring Call Filter* uses present/absent calls generated from the imaging software, which was designed to mimic the present/absent calls generated by GeneSpring software for Affymetrix data. The P (present), A (absent), M (marginal), and U (unknown) calls are based on the quality control flags listed in the CodeLink Call Filter.

- **(P) Present Call** - The spot receives a P call if it has a G quality control flag.
- **(A) Absent Call** - The spot receives an A call if it has an L quality control flag.
- (**M) Marginal Call** - The spot receives an M call if it has a C, I, or S quality control flag.
- **(U) Unknown Call** - The spot receives a U call if it has an M quality control flag.

Like the CodeLink Call Filter, you can specify the number of samples with either a P (Good) call or an A (Low) call needed to keep or remove the probe from further analysis. Samples that have an M or U call are not counted toward the number of samples needed for retention or elimination.

Median Variance Filter

Use the *Median Variance Filter* to retain probes with greater variation than the median variation gene. Genes whose expression does not vary are unlikely to be differentially expressed. Since it is likely that most genes are not differentially expressed or associated with a particular phenotype, the median variance across all genes is a reasonable model of null variation; i.e., the variation due to other factors. The variance (s^2) is calculated across all subjects for each gene. The null hypothesis is that these variances represent random and normally distributed noise. For each gene, the statistic $W = (N-1)s^2/median(s^2)$ is computed, where N is the number of observations of the gene. It is approximately chi-square distributed with $N-1$ degree of freedom [(Rosner, 2000), p246]. The system calculates a p-value for rejecting the null hypothesis and performs the False Discovery Rate (FDR) multiple testing correction (Benjamini & Hochberg 1995), setting the FDR to 10%. The result is a list of genes with significantly greater variation than the median variation across genes with, at most, 10% of that list including genes that have true variation less than, or equal to, median variation.

References

1. Fundamentals of Biostatistics. Bernard Rosner, 5th edition Duxbury Thomson Learning. Pacific Grove, CA USA
2. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Coefficient Variation Filter

Use the *Coefficient Variation Filter* to remove probes whose variation across samples is higher than a threshold. The Coefficient Variation filter measures the consistency of the probes across all samples. The coefficient of variation (CV) of each probe is calculated as standard deviation divided by mean. A high CV value reflects inconsistency among the samples within the group. For a two-group comparison study, the CV of each group is calculated independently. You can predefine a cutoff value, where genes with CVs above the cutoff value are removed. You can also define different cutoff thresholds to control the consistency levels of the genes.

Negative Control Filter

Use the *Negative Control Filter* to keep or remove probes based on whether the negative control probes are above or below detection limits. The negative control probe filter takes advantage of the negative controls on the CodeLink arrays to create thresholds for detection. Detection limits are calculated for each array individually based on the negative controls on that array only. The detection limit is equal to the mean plus two standard deviations of the set of negative probes with the highest mean. Both the mean and standard deviation are calculated based on a 5% trimmed data set.

Heritability Filter

Use the *Heritability Filter* to limit the probes analyzed to those with a high genetic heritability. This filter eliminates probes for transcripts whose environmental influence on expression is high compared to the strict genetic influence. Broad sense heritability has been calculated on the public HXB/BXH recombinant inbred rat panel normalized using cyclic LOESS. The broad sense heritability is calculated for each probe separately using an ANOVA model. Because the public data sets are based on the CodeLink Whole Genome Rat Array, this filter is only available for data sets on that chip. A minimum heritability threshold for inclusion must be specified. All filtering is done on the probe level.

eQTL/bQTL Filter

Another way to prioritize genes for analysis is to limit those that are considered to genes whose transcription levels are controlled from the same genetic region that controls the phenotype/behavior of interest (e.g., Tabakoff et al. 2009). We have identified expression quantitative trait loci (eQTL) for probes from brain tissue of the public HXB/BXH recombinant inbred rat panel normalized using cyclic LOESS. The user chooses a significance threshold for eQTL and the appropriate bQTL to compare to. Probes are retained if their eQTL is sig-

nificant and the location of the marker (SNP) with the maximum association with transcript expression is within the chosen bQTL limits. All filtering is done on the probe level.

Gene List Filter

Filtering by Gene List allows you to select a gene list that has been previously created and either keep or remove all the genes within that gene list.

Clustering Filtering Procedures

Variation Filter

Filter by Variation to reduce the number of probes considered for the clustering analysis by only retaining the genes with the largest variance across all samples (group is not accounted for in this calculation). You can indicate how many probes to retain by either specifying a percent or an exact number.

Fold Change Filter

Filter by Fold Change to reduce the number of probes considered for the clustering analysis by retaining probes with the largest difference between the maximum expression value for that probe and the minimum expression value for that probe. You can indicate how many probes to retain by either specifying a percent or an exact number.

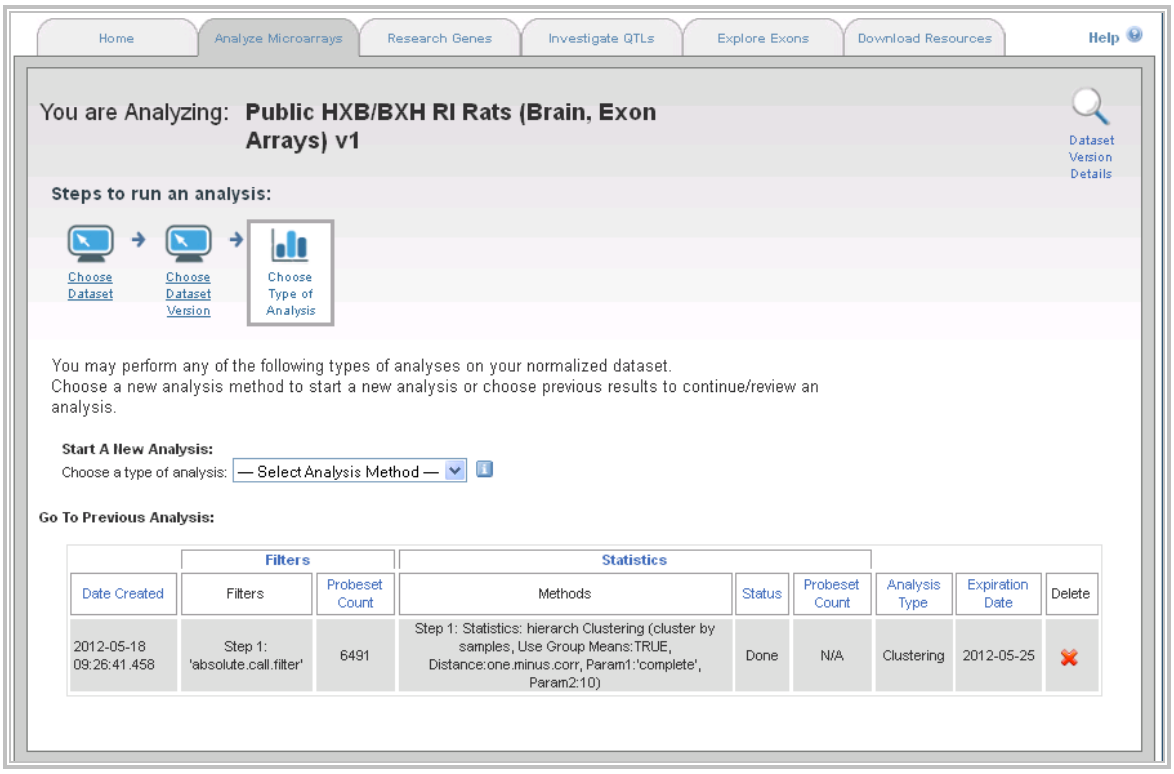
Types of Statistical Analysis

The PhenoGen website provides three different types of analyses that can be performed on your "in-silico" dataset:

- "Differential Expression Analysis" on page 68
- "Correlation Analysis" on page 70
- "Clustering Analysis" on page 72

Note: You must upload phenotype data for a dataset version if you want to run correlation analysis. See "Using Phenotype Data in Correlation Analysis" for details.

For Exons only, a *Previous Analysis* section displays if the dataset has been previously analyzed. Previous analyses expire after seven days.



Differential Expression Analysis

The ultimate goal of differential expression analysis is to select genes whose expression values are significantly different between two or more groups of samples. Statistical tools are available for the following types of analysis:

- Differential expression in two groups.
- Differential expression using one-way ANOVA.
- Differential expression using two-way ANOVA.
- Differential expression in replicate datasets.

The type of statistical tools available to the user are dependent on the number of groups specified during the Preparing Datasets process. If more complicated analyses are needed, you can download the raw CEL files or the normalized data for analysis using the statistical package of your choice. See "Downloading a Dataset".

Statistics for Differential Expression in Two Groups

If you have grouped the arrays in your dataset into only two groups, you can perform parametric, non-parametric, or 2-way ANOVA statistical analysis.

Parametric Analysis

The parametric analysis is a two-sample t-test assuming equal variances and is executed using the R function *t.test*. This test is considered parametric because it assumes that the distribution of expression values is normal within each group. However, expression data is often not distributed normally and there are often not enough observations to assume that the central limit theorem applies.

Non-Parametric Analysis

A non-parametric test does not depend on the distribution of the expression data. The Wilcoxon rank sum test is invoked when you specify a non-parametric test. This analysis is done in R using the function *wilcox_test* from the *coin* package. When the sample size is under 50, the exact p-value is used, otherwise a normal approximation is calculated. The parametric t-test is more powerful when the data is truly normally distributed, but the non-parametric test is robust for data that is not normally distributed.

Two-way ANOVA Analysis

Use a two-way ANOVA Analysis to input factors or use factors from the array attributes in a two-way analysis of variance. You can test for four different effects; main effect of factor 1, main effect of factor 2, the interaction effect between factor 1 and factor 2, or the overall model effect. To test for the interaction effect or the overall model effect, a regression model is used that contains the main effects for both factor 1 and factor 2 plus the interaction effect, a true two-way ANOVA. To test for main effects, the interaction effect is not included in the model. F-statistics are reported in all cases. Factors that are entered as character values are treated as categorical, and factors that are entered as numerical values are treated as continuous.

Statistics for Differential Expression with More than Two Groups

If you have grouped the arrays in your dataset into more than two groups, the following types of statistical analyses are available:

- One-way ANOVA
- Two-way ANOVA
- T-test with noise distribution

One-way ANOVA Analysis

Use a one-way ANOVA model to compare the within-group variance and the between-group variance of selected groups. You can choose from any of the possible pair-wise contrasts (e.g., Group1 vs. Group 2) or the overall effect of group (factor effect) on the model. When there are more than four groups, you can only use the factor effect. For the pair-wise contrasts, a moderated t-statistic is used to test significance, and for factor effects, a moderated F-statistic is used. For the moderated t-statistic and F-statistic, the standard deviation of the ordinary test is "shrunk" to reflect information that is borrowed across genes (Smyth 2004).

References

Smyth GK (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology 3(1), Article 3.

Two-way ANOVA Analysis

Use a two-way ANOVA Analysis to input factors or use factors from the array attributes in a two-way analysis of variance. You can test for four different effects; main effect of factor 1, main effect of factor 2, the interaction effect between factor 1 and factor 2, or the overall model effect. To test for the interaction effect or the overall model effect, a regression model is used that contains the main effects for both factor 1 and factor 2 plus the interaction effect, a true two-way ANOVA. To test for main effects, the interaction effect is not included in the model. F-statistics are reported in all cases. Factors that are entered as character values are treated as categorical, and factors that are entered as numerical values are treated as continuous.

T-Test with Noise Distribution Analysis

Use a t-test with noise distribution to determine a list of genes that are differentially expressed between two groups in a replicate experiment, i.e., "high" selected line from replicate 1, "low" selected line from replicate 1, "high" selected line from replicate 2, and "low" selected line from replicate 2. This method was first introduced by Eaves et al. in 2002.

This method involves first calculating a modified t-statistic for each probe(set) for each replicate separately where the traditional sample variances are replaced with a "pooled" variance. The pooled variance is calculated for each group using a weighted mean between the observed variance and a mean local variance. The weights are 2 to 1 where the larger of the two variances is given the larger weight. To calculate the mean local variance, the data is first sorted by mean expression for each probe(set), the mean local variance is then calculated as the mean of the variances of the 250 probe(set)s immediately below the probe(set) of interest and the 250 probe(set)s immediately above the probe(set) of interest.

After t-statistics are calculated for each probe(set) and each replicate, the probe(set)s are separated into two groups. Probe(set)s are placed in the null distribution if their t-statistics show opposite signs in the two replicate experiments. The t-statistics for these experiments are used to generate a null distribution of t-statistics for p-values to be based on. Instead of individual p-values being calculated for each probe(set), an initial p-value threshold is set and it is determined whether or not probe(set)s meet this criteria. Using this method, the type I error rate (p-value) is determined by the product of the following three probabilities:

1. Probability of a t-statistic greater than the one observed given the gene is not differentially expressed in replicate dataset1 (i.e., in null distribution).
2. Probability of a t-statistic greater than the one observed given the gene is not differentially expressed in replicate dataset2 (i.e., in null distribution).
3. Probability of having the two t-statistics showing the same direction of differential expression (i.e. 0.5).

Therefore, the percentiles from the null distribution used to determine 'significance' can be calculated for a specific error rate by taking the square root of the fraction, the probability of having the two t-statistics showing the same direction of differential expression divided by the specified error rate. A gene is considered significant if the observed t-statistic for each replicate is larger than the threshold t-statistic determined from the null distribution for that replicate. Exact p-values are not reported for this statistical method, and therefore, multiple testing corrections cannot be implemented.

References

Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, Todd JA, Glynne RJ (2002). Combining mouse congenic strain and microarray gene expression analyses to study a complex trait: The NOD model of type1 diabetes. Genome Research 12:232-243.

Correlation Analysis

A correlation analysis is used to determine if two variables are associated with each other. The Correlation Analysis tool on the PhenoGen website searches for genes whose expression value correlates with a phenotype value. You can use the correlation analysis tools with one of your own experiments or with one of the public datasets on the website (see "Public Datasets" on page 14 for details). You upload phenotype data which can be any continuous measurement.

For example, if you are interested in genes expressed in the brain that are correlated with alcohol preference in mice, you can upload a file containing alcohol preference values for any or all of the inbred or recombinant inbred strains for which whole brain gene expression data is available on the PhenoGen website see "Public Datasets" on page 14 for details). A positive correlation between alcohol preference and the expression value of a transcript in strains where the expression for that transcript is high might indicate a relationship between that gene and alcohol preference. Similarly, a negative correlation might indicate an inverse relationship.

Statistics for Correlation Analysis

For correlation analysis, mean expression values are calculated within strain. These mean values are correlated with the strain phenotypic measures that you upload. You can choose either a Pearson Correlation Coefficient or a Spearman Rank Correlation Coefficient to calculate correlation.

Pearson Test

The Pearson correlation coefficient is a parametric test for a linear relationship between two variables. A parametric test is more powerful when the two variables involved are truly normally distributed, but a non-parametric test is more accurate when this normality assumption is not met.

Spearman Test

The Spearman rank correlation is a non-parametric test that looks at the correlation of the ranks of the values rather than the actual values. A non-parametric test is necessary when the distribution of either of the variables is not normal.

Multiple Testing Adjustment

There are eleven options for adjusting p-values for multiple testing. Multiple testing adjustments are only used for statistical analysis when doing a differential expression or a correlation analysis. They are split into four categories:

- False Discovery Rate (FDR)
- General
- Permutation
- None

False Discovery Rate

1. Benjamini and Hochberg (BH)
2. Benjamini and Yekutieli (BY)
3. Storey

References

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. JR Statist Soc B 57:289-300.

Benjamini Y and Yekutieli D (2001). The control of the false discovery rate in multiple hypothesis testing under dependency. Annals of Statistics 29(4):1165-1188.

Storey JD (2002). A direct approach to false discovery rates. JR Statist Soc B 64:479-498.

General

4. Bonferroni
5. Holm
6. Hochberg
7. Sidak Single Step (SidakSS)
8. Sidak Step-Down (SidakSD)

References

Dudoit S, Shaffer JP, Boldrick JC (2002). Multiple hypothesis testing in microarray experiments. UC Berkeley Division of Biostatistics Working Paper Series, paper 110. http://www.bepress.com/ucbbiostat/paper110

Permutation

9. minP with permutation (only available for two groups)
10. maxT with permutation (only available for two groups)

References

Westfall PH and Young SS (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, John Wiley & Sons.

None

11. No Test

All multiple testing adjustments are conducted in R. Adjustments 1, 2 and 4 through 8 are done using the R function *mt.rawp2adjp*. Adjustment 9 uses *mt.maxT* and adjustment 10 uses *mt.minP*. Adjustment 3 uses *qvalue*.

For several methods, only the threshold value for α needs to be specified to complete the analysis. For Adjustment 3, the method for estimating the tuning parameter must be specified in addition to the alpha threshold. You can use a "smoother" approach or a "bootstrap" approach. For most cases, the "smoother" approach works better, but there are situations when it will fail. Therefore, "bootstrap" is overall a safer option.

With the two methods of adjustment that require permutation (9 and 10), two more parameters must be specified because these two methods require that the statistical analysis be done for each permutation. The two additional parameters are:

- The type of test to be conducted. The choice here overrides the results from the statistical analysis specified previously.
- The number of permutations to use in estimating the p-value.

Note: The last method, 'No Test', provides the option of not using any of the multiple comparison adjustments.

Clustering Analysis

The PhenoGen website provides the ability to cluster gene expression values by samples, groups, probesets, or when using hierarchical clustering, by both samples or groups and probe sets. When you perform clustering on a dataset, you have additional options for filtering the probes that are included in the analysis.

In addition to the filtering options available for other types of analysis, the following options are also available for clustering:

- Variation
- Fold Change

See "Clustering Filtering Procedures" for details.

Statistics for Cluster Analysis

After you filter, you must specify the clustering algorithm to use, the expression values to use, which object is to be clustered, the distance measure to use, how many clusters you want to use, and which dissimilarity measure to use (for hierarchical clustering).

- **Clustering Algorithm** - Choose whether to cluster using the **hierarchical** or **k-means** algorithm.
- **Mean Expression Values** - Choose whether to cluster using individual expression values for each sample in your dataset or, if your dataset has more than two groups, to use the mean expression values for each group.
- **Cluster Object** - Choose whether to cluster samples, groups, or probes. When you choose hierarchical clustering, you can also cluster by both samples or groups and probes to get a heat map representation of the data.
- **Dissimilarity Measure** - Choose the method for determining dissimilarity between clusters.
- **Distance Measure** - Both clustering algorithms are based on the distance that one cluster object is from the other, in other words, the dissimilarity between objects. In the PhenoGen website, distance (dissimilarity) can be calculated using two different measures; Euclidean distance or one minus the

correlation. The Euclidean distance is the square root of the sum of squared differences. In general, one minus the correlation is more commonly used in microarray analyses because it is both location and scale invariant.

Hierarchical Clustering Method

The *hierarchical clustering method* on the PhenoGen website uses bottom-up methodology where each cluster object starts off as its own cluster. Next, the two clusters that are the most similar are combined into one cluster. This process is repeated until all clusters have been combined to form one cluster that contains all cluster objects. When using hierarchical clustering, you must choose a between-cluster dissimilarity measure.

Hierarchical clustering is implemented using the *hclust* function in R. You can also specify the number of clusters to form. Generally in hierarchical clustering, the number of clusters does not need to be specified a priori, but on the PhenoGen website, when you specify the number of clusters to form, cluster objects can be placed into groups, and if the cluster objects are probes, then these individual groups can be downloaded as gene lists for further exploration and analysis. In addition to a numerical representation of the clusters, a dendrogram is also created and displayed when a single cluster object is chosen. When you choose to cluster on both samples (or groups) and probes, you do not have to specify the number of clusters, and the only output generated is a heat map with samples or groups along the x axis and probes along the y axis. In the heat map, the expression intensity values are represented as a z-score calculated using the mean and standard deviation for that particular probe. The z-score is represented on a color scale from bright red to bright green where red indicates a larger negative z-score value, and green indicates a larger positive z-score. See "Viewing Heat Maps" for details.

Dissimilarity Measure

The distance between clusters is calculated using the Lance Williams dissimilarity update formula according to the measure you chose. There are four different options for the between-cluster dissimilarity measure:

- **Single** - Minimum difference between points in different clusters.
- **Complete** - Maximum difference between points in different clusters
- **Average** - Average of all distances between points in different clusters.
- **Centroid** - Difference between cluster centroids.

K-Means Partitioning Method

The other clustering algorithm, the *k-means partitioning method*, iteratively updates the cluster centers until the sum of squared distances from each observation to its cluster center is minimized. The Hartigan and Wong method as implemented in the function *kmeans* in R is used for the k-means analysis. This algorithm requires that initial estimates of cluster centers be given. Currently, cluster objects are chosen at random to use as starting locations. Since different starting locations might generate different results, you should be aware that if the same analysis is carried out again on the website, the results may differ.

References

A partial list of references for additional information about the clustering methods follows.

1. Speed, T (2003). Statistical Analysis of Gene Expression Microarray Data. New York: Chapman and Hall/CRC.

References for 'hclust' in R

1. Everitt, B. (1974). Cluster Analysis. London: Heinemann Educ. Books.
2. Hartigan, J. A. (1975). Clustering Algorithms. New York: Wiley.
3. Sneath, P. H. A. and R. R. Sokal (1973). Numerical Taxonomy. San Francisco: Freeman.
4. Anderberg, M. R. (1973). Cluster Analysis for Applications. Academic Press: New York.

5. Gordon, A. D. (1999). Classification. Second Edition. London: Chapman and Hall / CRC
6. Murtagh, F. (1985). "Multidimensional Clustering Algorithms", in COMPSTAT Lectures 4. Wuerzburg: Physica-Verlag (for algorithmic details of algorithms used).
7. McQuitty, L.L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. Educational and Psychological Measurement, 26, 825–831.

References for 'kmeans' in R

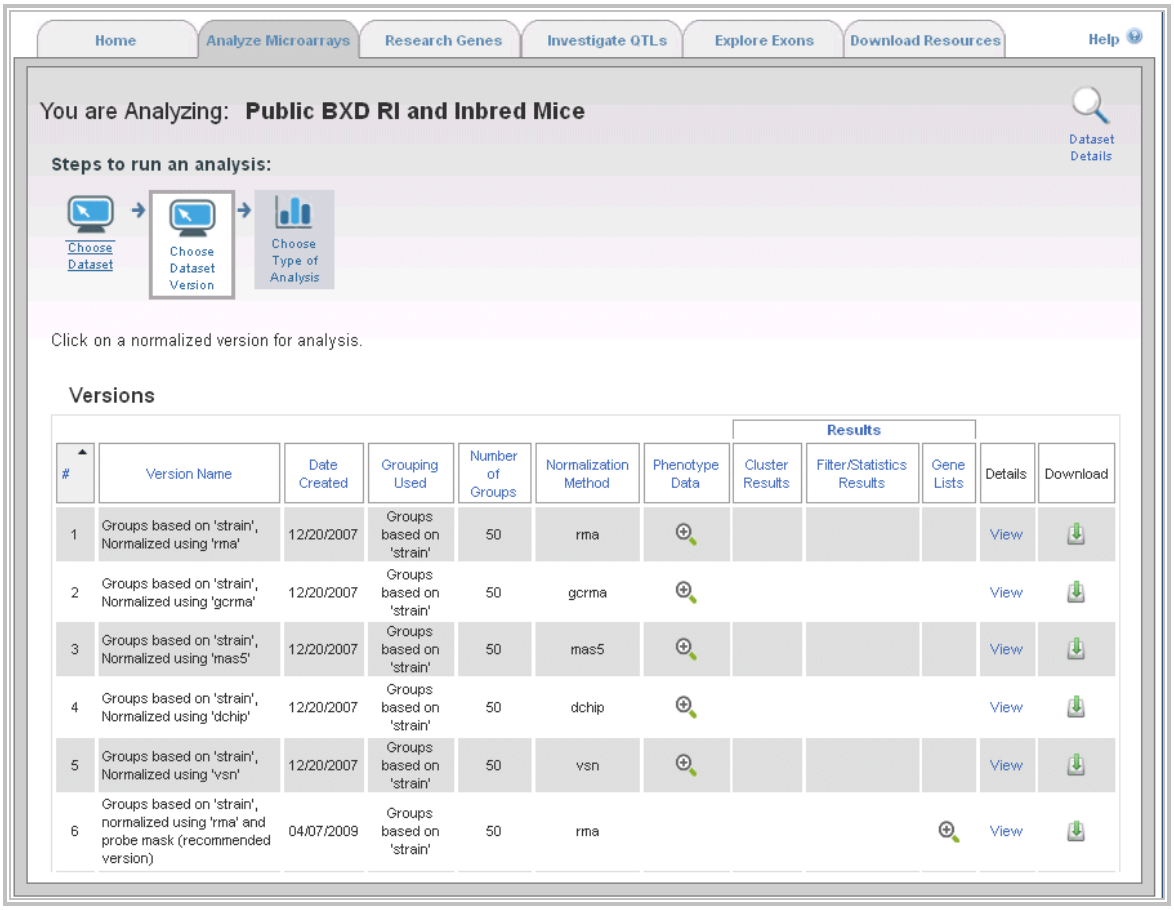
1. Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21, 768–769.
2. Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100–108.
3. Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory 28, 128–137.
4. MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman, 1, pp. 281–297. Berkeley, CA: University of California Press.

Filtering and Analyzing Datasets

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.
3. Click the dataset you want to analyze.

   Note: Datasets that are ready for analysis have a checkmark in the **QC Complete** and the **Arrays Grouped and Normalized** columns.

4. Click the Normalized version you want to analyze. Click **View** in the **Details** column to view the parameters used for normalizing the version.



5. Choose a type of analysis:
   - "Differential Expression Analysis" on page 68 - Proceed to step 7.
   - "Correlation Analysis" on page 70 - Proceed to step 6.
   - "Clustering Analysis" on page 72 - Proceed to step 7.
6. Do one of the following:
   1. Click the **Create New Phenotype** link at the top right, and choose an option:
      - **Upload Phenotype Data -** Allows you to upload a phenotype data file.
      - **Enter New Phenotype Data** - Allows you to enter new phenotype data.
      - **Copy Existing Phenotype Data** - Allows you to copy and edit existing phenotype data. If phenotype data does not exist for the version, this option is not available.

      Note: See "Using Phenotype Data in Correlation Analysis" for upload, enter, copy, and delete instructions, then proceed to step 7.

      OR
   2. Choose a **phenotype** from the *Phenotype Values* table, then proceed to step 7.
7. Choose a gene filtering method.

   Filtering Affymetrix Arrays

   If your experiment contains Affymetrix arrays, choose:

- **Affy Control Genes** - Proceed to step 8.
- **MAS5 Absolute Call Filter** - Specify whether to **Keep** or **Remove** probes and the values for groups, then proceed to step 8.
- **Heritability Filter** - Specify the **panel** to use, enter the **Minimum Heritability Criteria**, then proceed to step 8.
- **ebQTL/eQTL** - Choose a QTL list, then proceed to step 8.
- **Gene List Filter** - Specify whether to **Keep** or **Remove** probes for a selected gene list. This is a good option when you want to review results from a previous analysis. Proceed to step 8.

Filtering CodeLink Arrays

If your experiment contains CodeLink arrays, choose:

- **CodeLink Control Genes Filter** - Proceed to step 8.
- **CodeLink Call Filter** - Specify whether to **Keep** or **Remove** probes and the values for groups, then proceed to step 8.
- **GeneSpring Call Filter** - Specify whether to **Keep** or **Remove** probes and the values for groups, then proceed to step 8.
- **Median Filter** - Specify the **Filter threshold**, then proceed to step 8.
- **Coefficient Variation Filter** - Specify **and** or **or** and the values for groups, then proceed to step 8.
- **Negative Control Filter** - Specify whether to **Keep** or **Remove** probes and the values for groups, specify the Trim percentage, then proceed to step 8.
- **Heritability Filter** - Specify the **Minimum Heritability Criteria** then proceed to step 8.
- **bQTL/eQTL** - Choose a QTL list, then proceed to step 8.
- **Gene List Filter** - Specify whether to **Keep** or **Remove** probes for a selected gene list. This is a good option when you want to review results from a previous analysis. Proceed to step 8.

Additional Options for Cluster Analysis

If you are doing a cluster analysis, the following additional options are available:

- **Variation Filter** - Specify how many probes to retain by either entering a percent or an exact number, then proceed to step 8.
- **Fold Change Filter** - Specify how many probes to retain by either entering a percent or an exact number, then proceed to step 8.

8. Click **Run Filter** to run the filter. After the filter runs, you can choose another filter to refine your filtering criteria, and click **Run Filter** again.
9. Click **Next** to proceed to statistical analysis or clustering when you are satisfied with the number of probes remaining.
10. Select the type of statistics test you want to run:

Differential Expression

- Parametric or non-parametric
- 1-Way ANOVA, then select the 1-Way ANOVA Parameter.
- 2-Way ANOVA, then select the P-value of Interest and choose the 2-Way ANOVA Factors.

Correlation Analysis

- Pearson or Spearman

Clustering

- Hierarchical, then select the use group means or individual sample values, the distance measure, the cluster object, the between cluster dissimilarity measure, and the number of clusters to report.
- K-means Partitioning, then select the use group means or individual sample values, the distance measure, the cluster object, and the number of clusters to report.

11. Click **Run Test**. You can run statistics multiple times with different options until you are satisfied with the results. Proceed to Step 12 for Differential Expression or Correlation. Proceed to Step 16 for Clustering Analysis.
12. Click **Next** to correct for **Multiple Testing**, if applicable. See "Multiple Testing Adjustment" for information about each selection in the drop-down list.
13. Based on your selection in step 12, you may need to:
    - Enter an **alpha level threshold**.
    - Select an **alpha or multiple correction threshold** from a drop-down list.
    - Enter an **alpha level threshold**, select the **type of test for permutation**, and select the **number of permutations**.
    - Enter the clustering parameters.

    Note: You can only cluster on probes if you have less than 5000 probes available.
14. Click **Run Adjustment**. You can run adjustments multiple times with different parameters until you are satisfied with the results.
15. Click **Next**.

Differential and Correlation Analysis

If the number of statistically significant genes is greater than zero (0) for a differential expression or correlation analysis, you can save the gene list.

16. Enter a **name** for the gene list.
17. Enter a **description** for the gene list.
18. **OPTIONAL:** Click the **Set description to the parameters used** checkbox to automatically populate the description field with the filtering and statistical analysis parameters you used.
19. Click **Save Gene List**. The gene list is saved, and a confirmation message displays. Click **Close**.

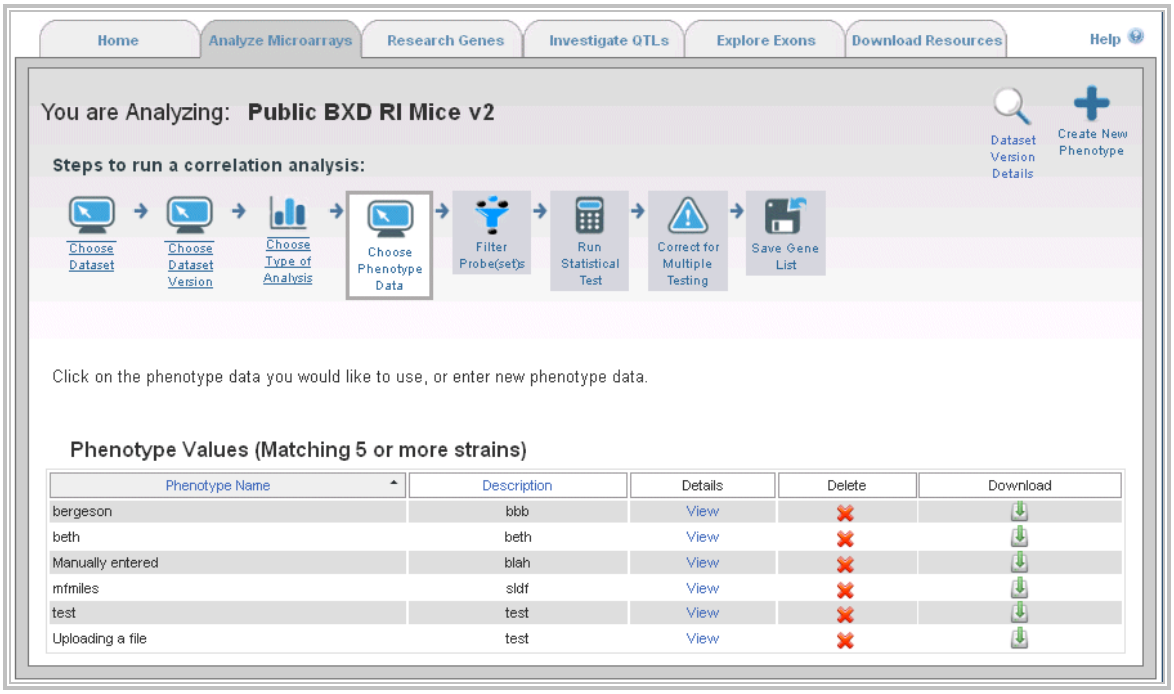
Clustering Analysis

When you perform a cluster analysis, the results display in a table below the *Analysis Parameters* table. You can review and save them. See "Saving Cluster Analysis Results" for details.

16. Click the **View Dendrogram** link beside the *Cluster Results* title to see the dendrogram.
17. Click **Save Results** to save the cluster results as a gene list. A confirmation message displays. Click **Close**.

Using Phenotype Data in Correlation Analysis

If you choose *Correlation Analysis* when you filter and analyze datasets, a page where you can run correlation analysis displays.



You can

- Click **Create New Phenotype** at the top right. You have the option of:
  - "Uploading Phenotype Data" on page 79.
  - "Entering Phenotype Data" on page 80.
  - "Copying Phenotype Data" on page 81.
- Click the **Delete** icon beside the phenotype values you want to delete.
- Click **View** to see a listing and graph of the phenotype values.

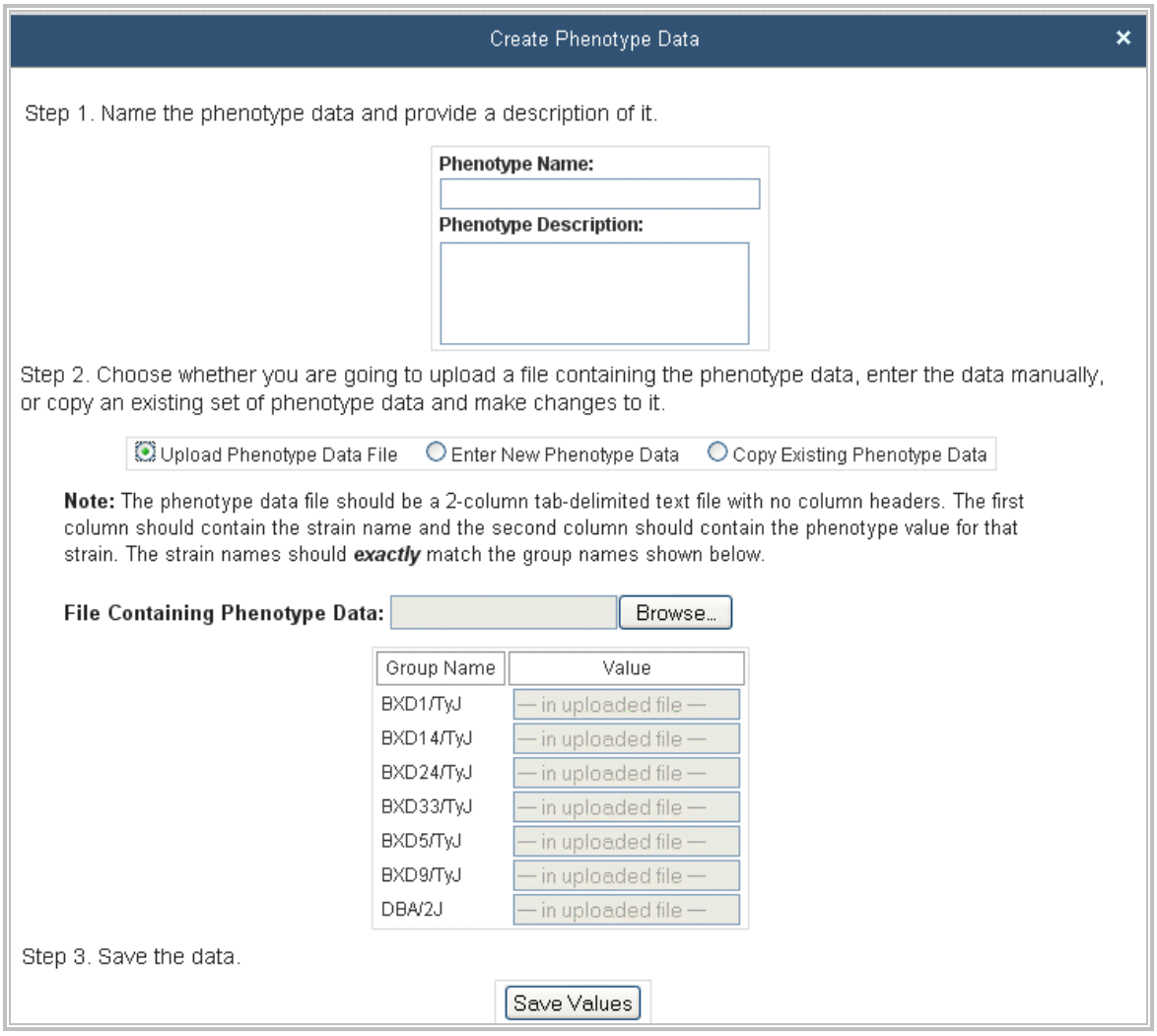
You can also re-normalize a public dataset during the Correlation Analysis. See "Re-normalizing a Public Dataset".

See "Filtering and Analyzing Datasets" for instructions to open the *Phenotypes* page.

Uploading Phenotype Data

Upload phenotype data from a file. The required format for a phenotype data file is a 2-column, tab-delimited text file with no column headers. The first column should contain the strain name, and the second column should contain the phenotype value for that strain. The strain names should exactly match the strain names in your dataset.

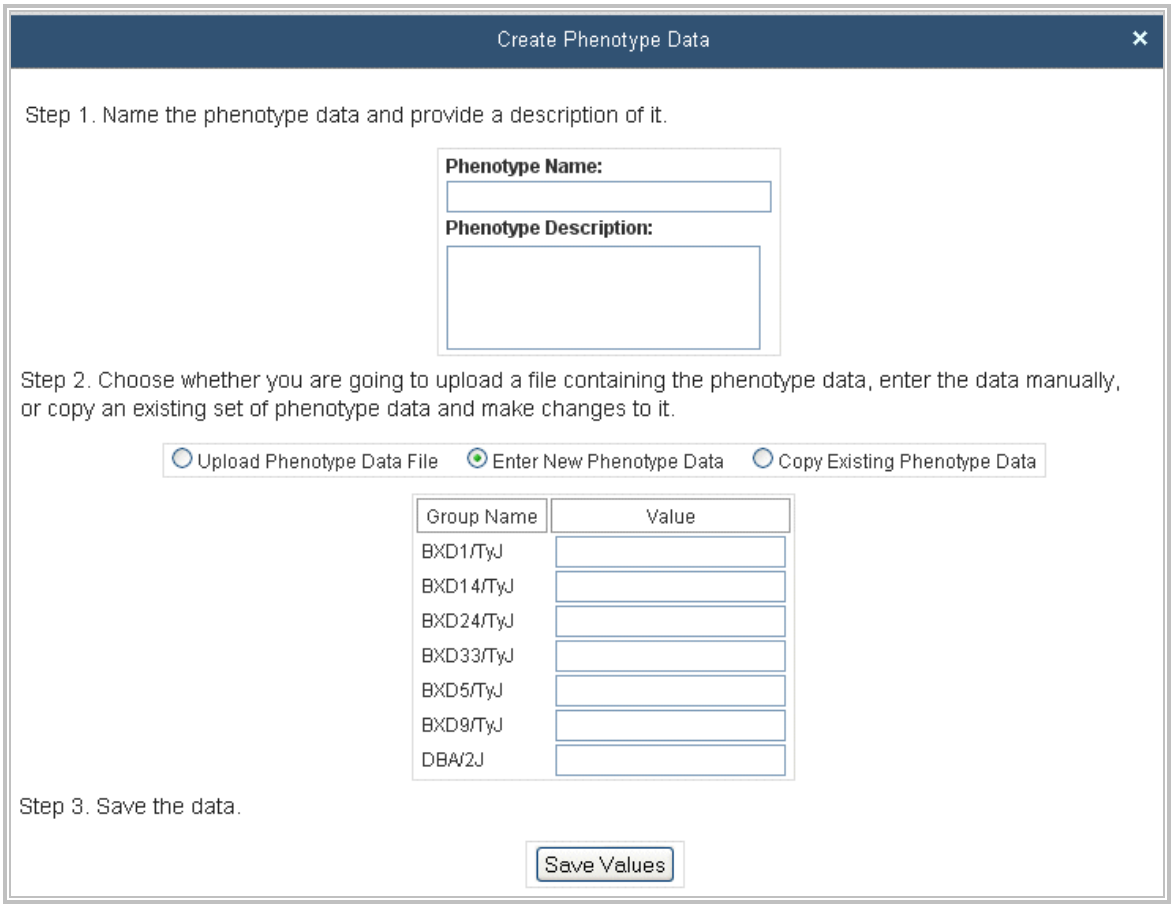
1. Click **Create New Phenotype** at the top right.
2. Enter a **name** for the phenotype.
3. Enter a **description** of the phenotype.
4. Select the **Upload Phenotype Data File** option button.
5. Click **Browse** to select the file that contains the phenotype data.
6. Click **Save Values**. A *Success* page displays with the number of matching strains uploaded.



Note: The strain names in your phenotype data file must match the strain names in the selected dataset exactly.

Entering Phenotype Data

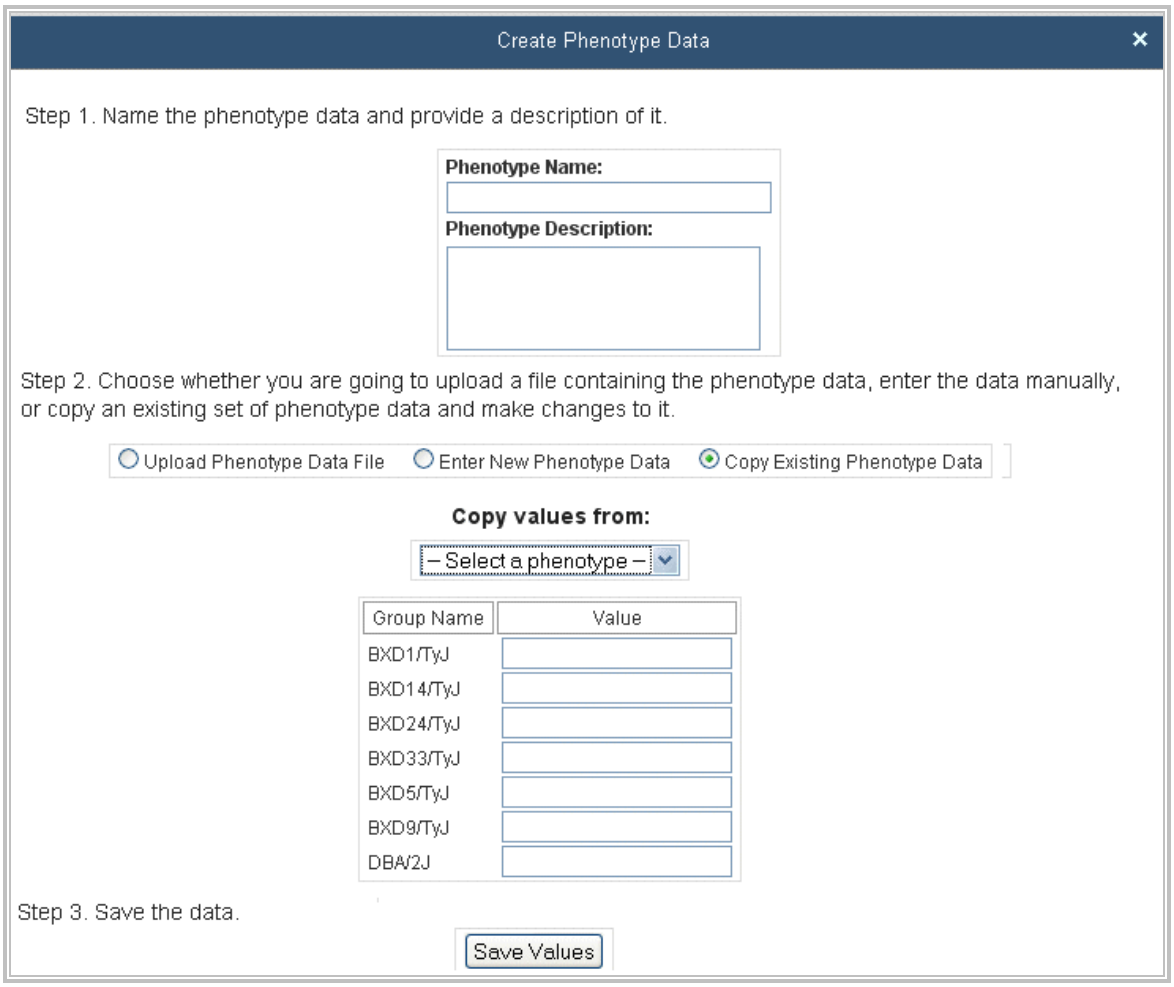
1. Click **Create New Phenotype** at the top right.
2. Enter a **name** for the phenotype.
3. Enter a **description** of the phenotype.
4. Select the **Enter New Phenotype Data** option button.
5. Enter values for each group in the list.
6. Click **Save Values**. A *Success* page displays.



Copying Phenotype Data

Copy, edit, and save phenotype data. This option is only available if Phenotype Values exist.

1. Click **Create New Phenotype** at the top right.
2. Enter a **name** for the phenotype.
3. Enter a **description** of the phenotype.

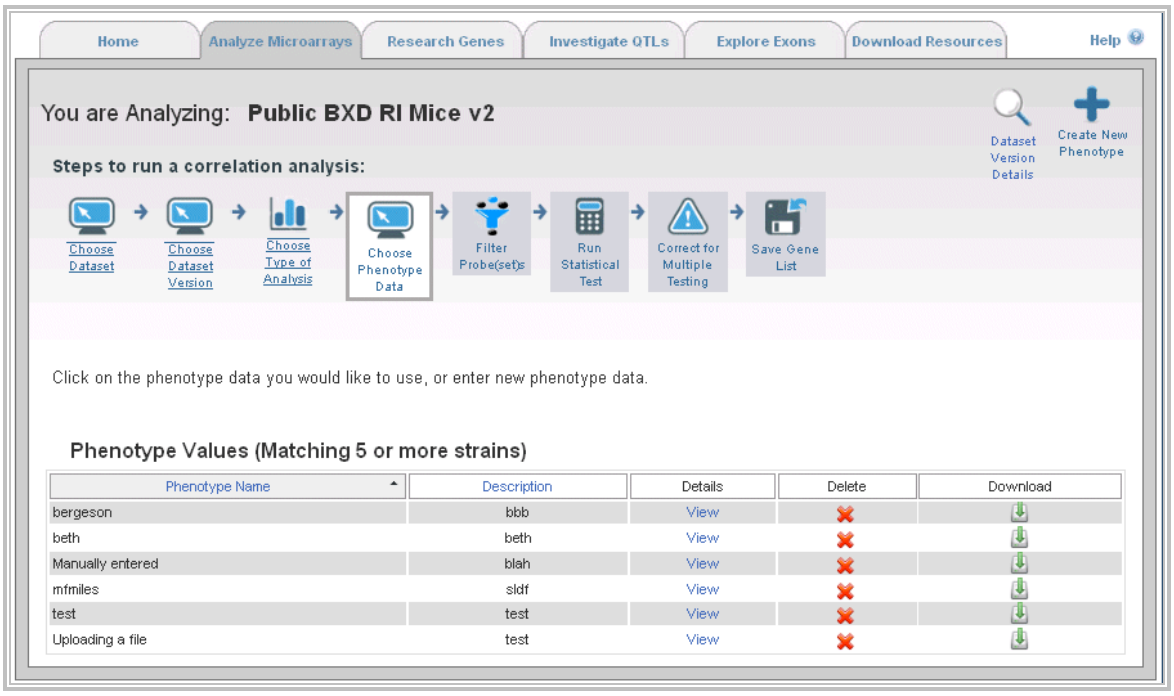


4. Select the **Copy Existing Phenotype Data** option button.
5. Choose an existing phenotype from the **Copy values from** drop-down list.
6. Enter or modify values for each group in the list.
7. Click **Save Values**. A *Success* page displays.

Re-normalizing a Public Dataset

When you run a correlation analysis on a dataset, you may have the option to re-normalize it, using only the strains for which you have entered phenotype data.

1. Choose a phenotype from the *Phenotype Values* table.

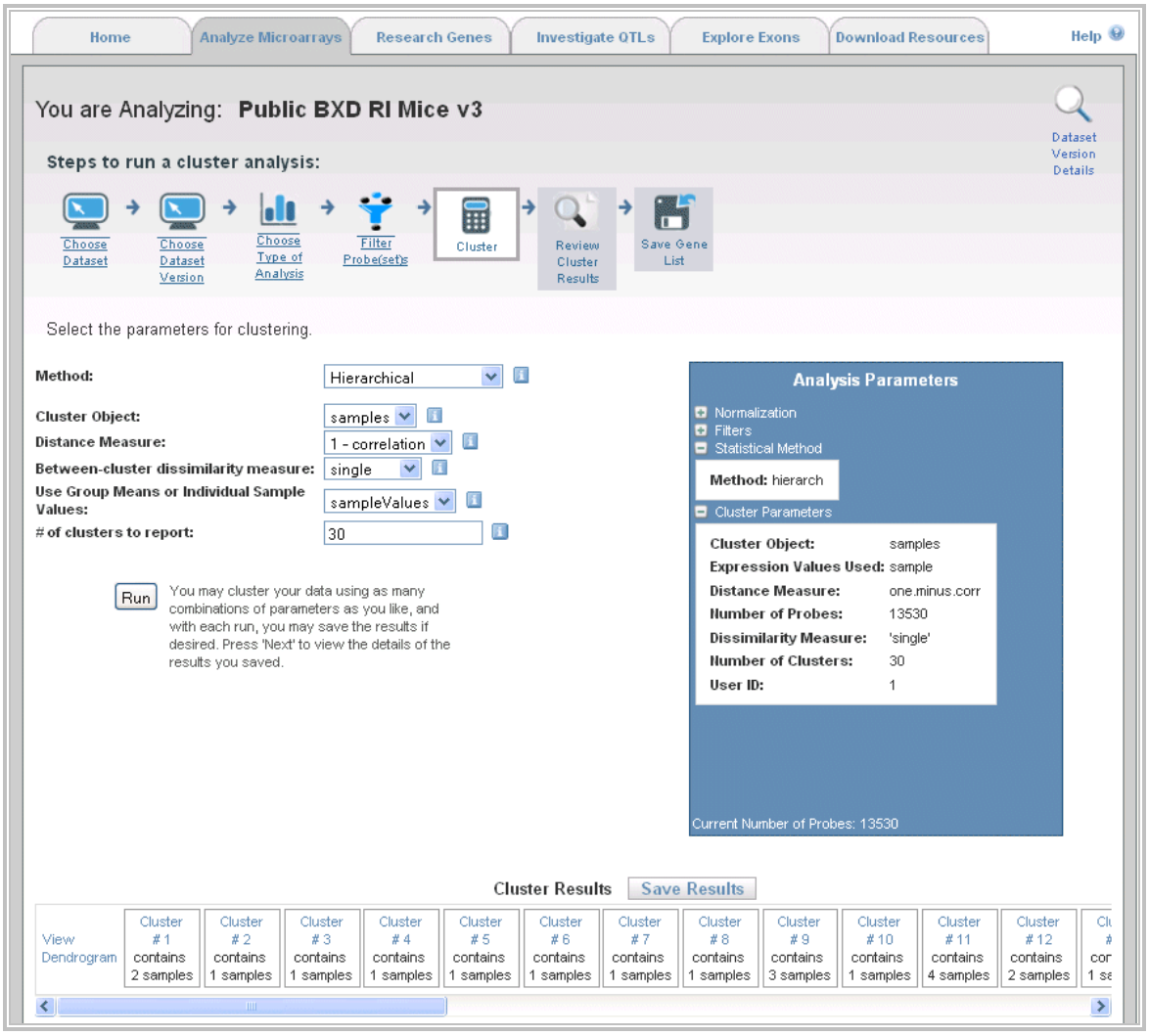


2. Click **Create New Phenotype** at the top right.
3. Enter a **name** for the phenotype.
4. Enter a **description** of the phenotype.
5. Select the **Enter New Phenotype Data** option button.
6. Enter values for the groups in the list.
7. Click **Save Values**. If you do not enter a value for each group in the list, a pop-up displays to ask if you want to re-normalize.

Saving Cluster Analysis Results

After you complete Cluster Analysis, you can save your results to view later.

1. Perform a cluster analysis. See "Filtering and Analyzing Datasets" for details.
2. Click **Save Results** when the analysis is complete.

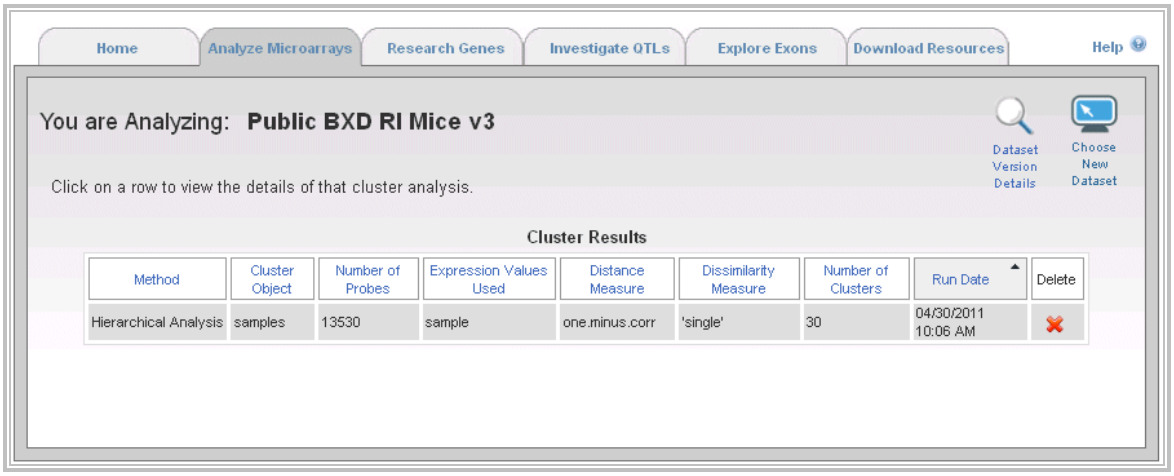


3. Click **Next** to view your results now. See "Viewing Cluster Analysis Results" for instructions on viewing your results later.

Viewing Cluster Analysis Results

You can view Cluster Analysis results for all the clustering analyses that you save. See "Saving Cluster Analysis Results".

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.
3. Click the **dataset** for which you want to see cluster results.
4. Click the **version** that contains the cluster results.
5. Click the **Magnify** icon in the Cluster column. The cluster results display.

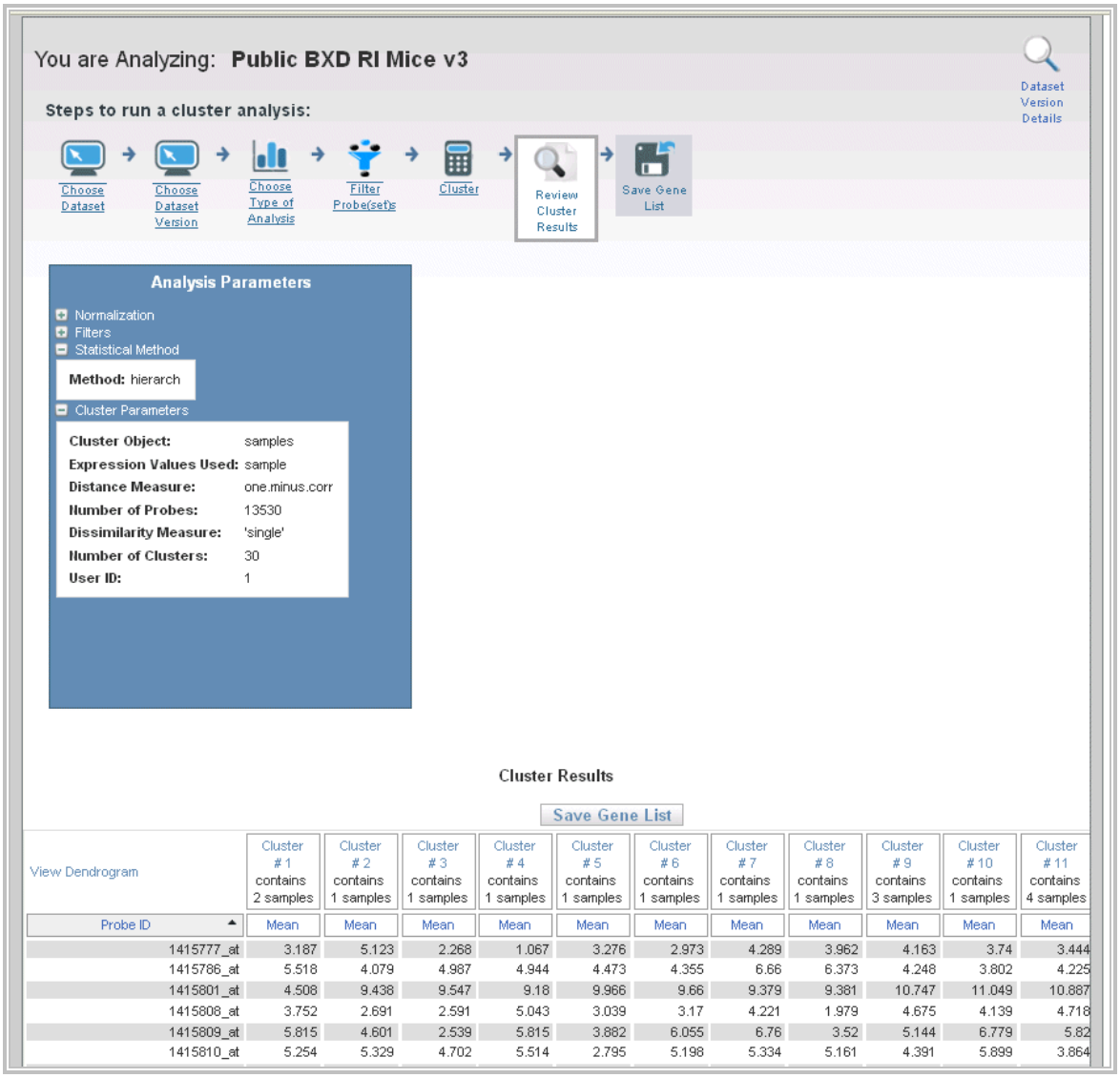


You can:

- Click a **Cluster Description** to view the results of that cluster analysis.
- Click the **Delete** icon beside the cluster results you want to delete.

Cluster Results

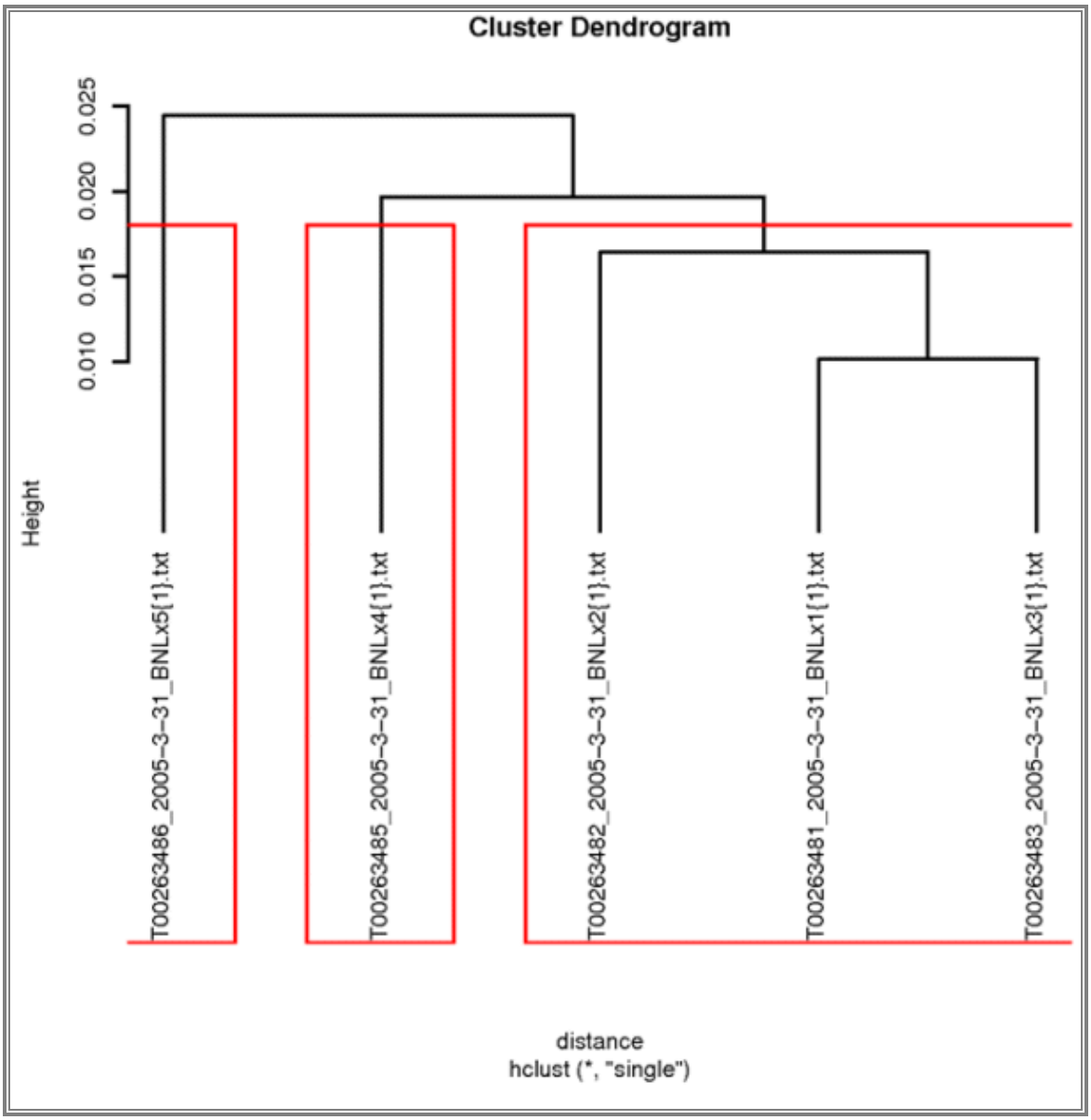
The *Cluster Results* page for the specific dataset displays the parameters used in the cluster analysis.



Viewing Dendrograms

A *dendrogram* provides a visual of the similarity between cluster objects. Along the horizontal axis, all cluster objects are listed. The cluster objects are "joined" at different levels until all cluster objects have been joined together into one group. The similarity between cluster objects or groups of cluster objects is related to the height at which they are joined. Cluster objects or groups of cluster objects that are joined at a lower height are more similar than cluster objects or groups of cluster objects that are joined at larger heights. The red boxes in the dendrogram indicate the groups of cluster objects that are defined at a specific height.

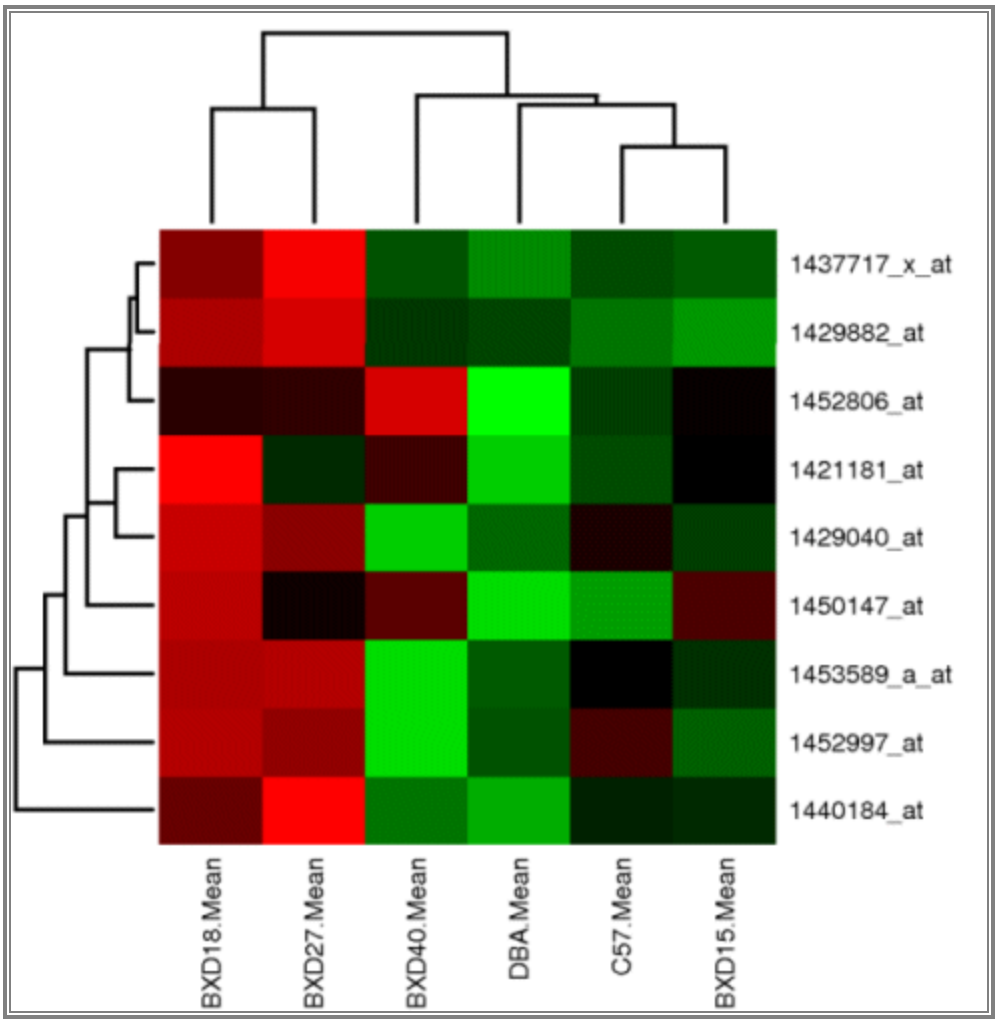
Depending on the type of cluster analysis, a dendrogram of the results may be available. If so, a **View Dendrogram** link displays.



Viewing Heat Maps

A *heat map* displays results when both probes and samples are clustered. It provides a visual representation of expression levels across all samples and all probes. Probes are represented along the vertical axis and samples are represented along the horizontal axis. The order of the probes and the samples depend on hierarchical clustering of the individual dendrograms along the left side and the top of the heat map. The expression intensity values are converted to a z-score that is calculated using the mean and standard deviation for that particular probe(set). The z-score is represented on a color scale from bright red to bright green where red indicates a larger negative z-score value and green indicated a larger positive z-score.

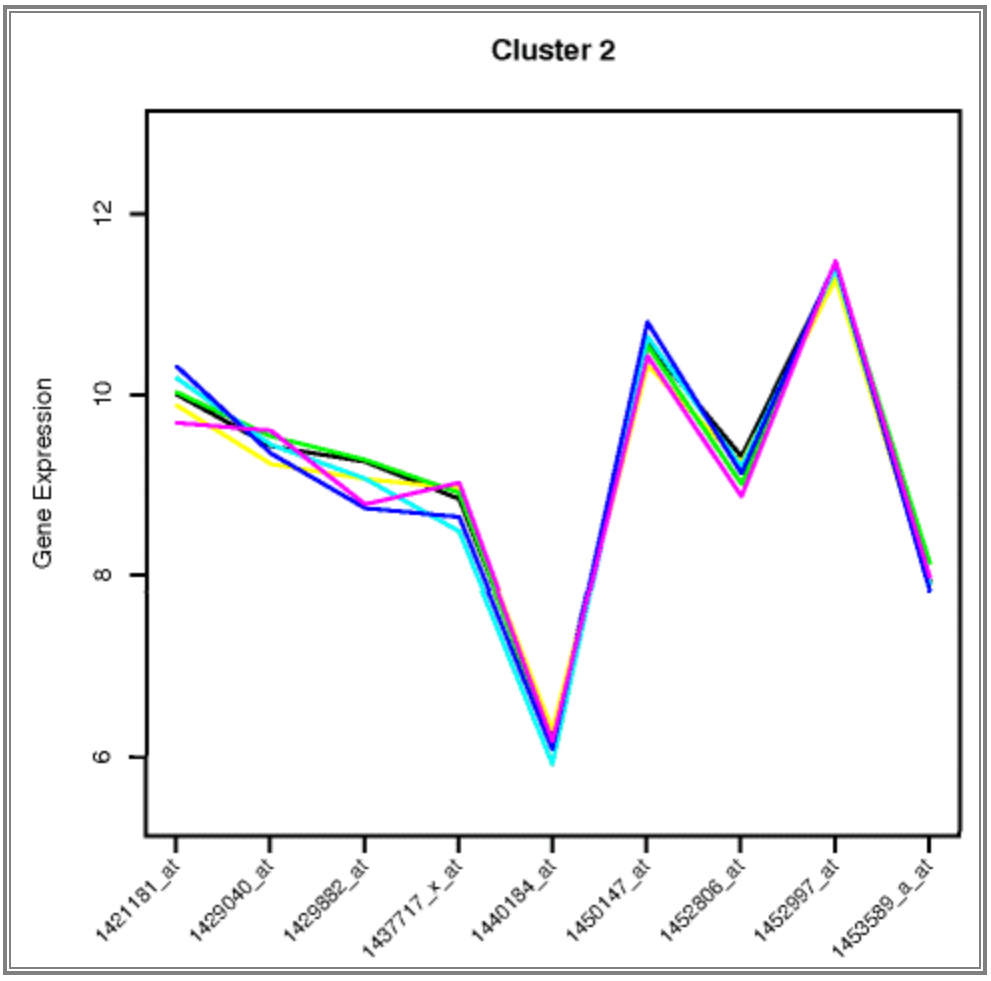
Depending on the type of cluster analysis, a heat map of the results may be available. If so, a **View Heat Map** link displays.



Viewing K-Means Graphs

When *k-means* partitioning is used, a graph is generated for each cluster. The graph has the non-cluster objects (i.e., if the cluster objects are probes then the non-cluster objects are samples) along the horizontal axis and gene expression is represented on the vertical axis. Each cluster object within a particular cluster is represented by a different color of line in the graph.

Depending on the type of cluster analysis, a k-means graph may be available. If so, a **View Graph** link displays.

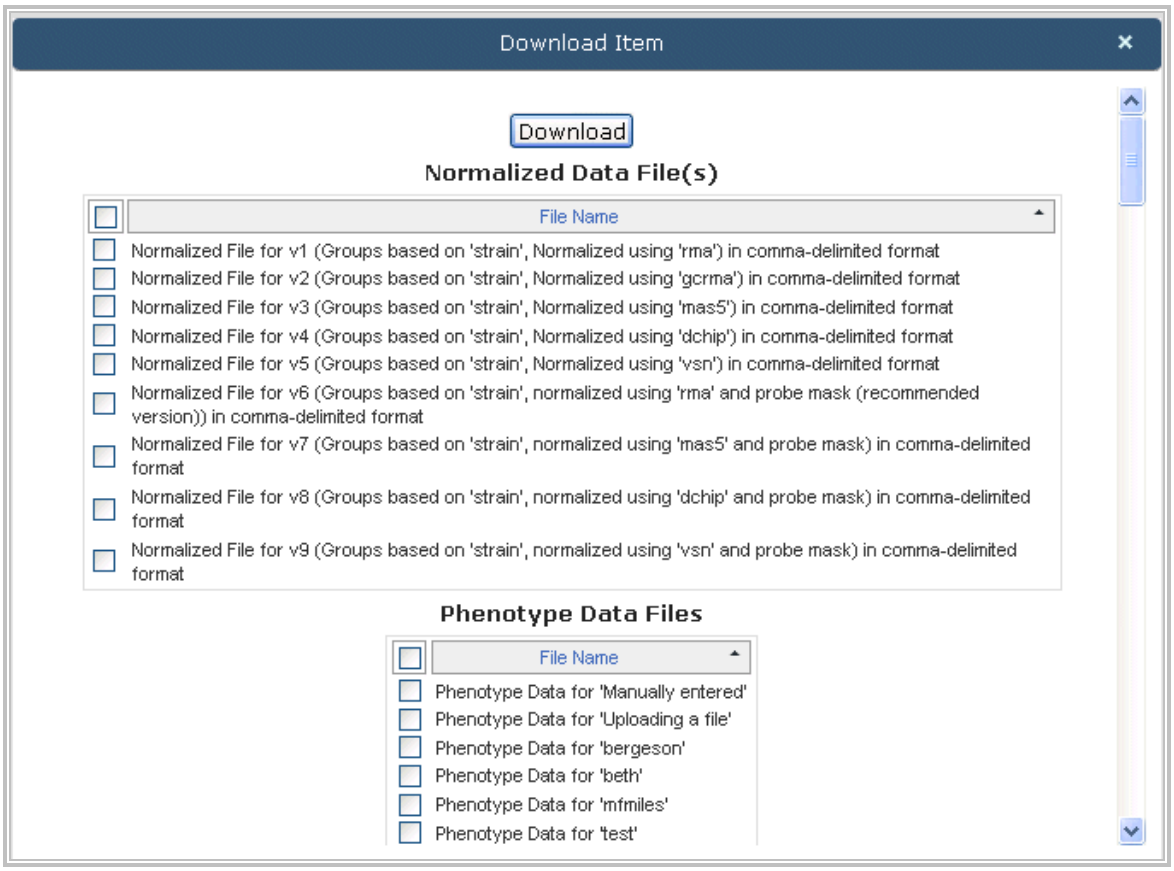


Downloading a Dataset

You can download raw data files and normalized versions of a dataset.

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.
3. Click the **Download** icon beside the dataset you want to download. The normalized data files, phenotype data files, and raw data files for each version of the dataset display.

   Note: If there are no normalized data files, only raw data files display.
4. Click the checkbox(es) next to the data files you want to download. Click the checkbox at the top of a column to select or deselect all.



5. Click **Download**.
6. Choose to open or save the files, and follow the instructions that display. These instructions vary depending on your Internet browser (e.g., Internet Explorer, Firefox, Safari, etc.). If files are large, an email that contains a link for downloading the prepared files is sent to you when they are ready to download.

Deleting Datasets and Versions

Deleting a Dataset

You can delete a dataset and all of its versions, if necessary.

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.
3. Click the **Delete** icon beside the dataset you want to delete.
4. Click the **Delete** button. A confirmation message displays. Click **Close**.

Deleting a Version

You can delete a version, if necessary.

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* page displays.
2. Click **Analyze microarray data** in the *What would you like to do* section. A page displays your datasets.
3. Click a dataset that has been normalized.

4. Click the **Delete** icon beside the version you want to delete.
5. Click **Delete**. A confirmation message displays. Click **Close**.

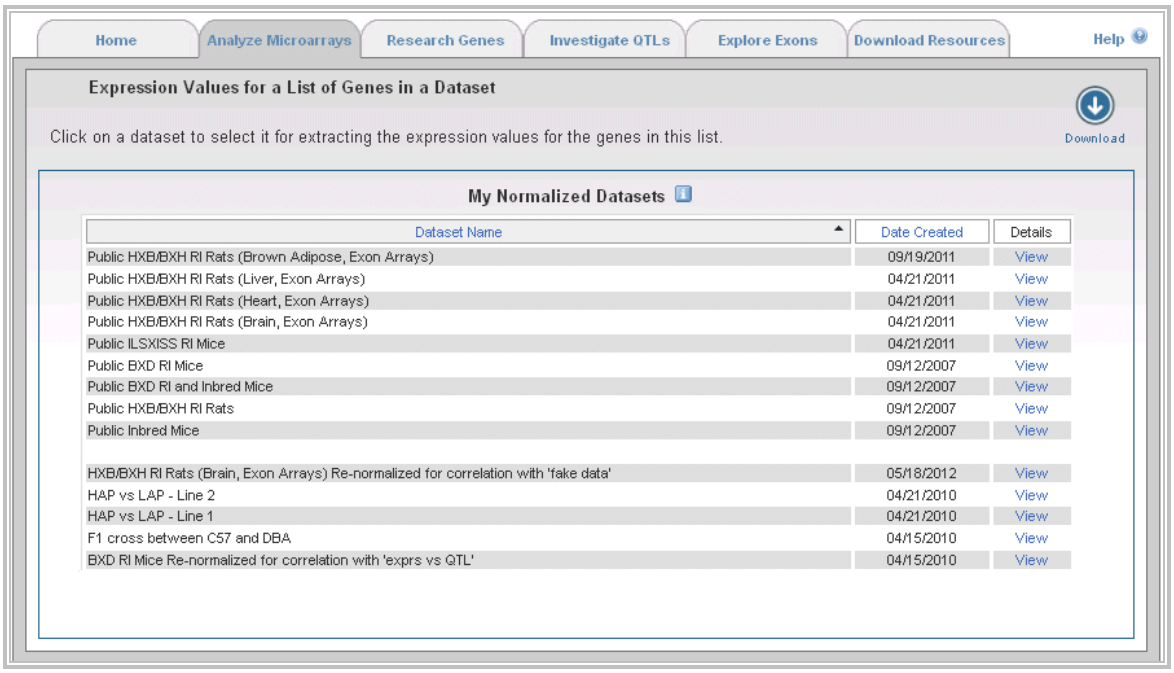
Viewing Gene Expression Data

The *View Gene Expression* page allows you to obtain gene expression intensity values for a gene or gene list from a normalized dataset, or specify a pre-created gene list. The results table displays group-level information for each gene, such as group means and group standard error.

You can view gene expression data from the Analyze Microarrays or the Research Genes tabs:

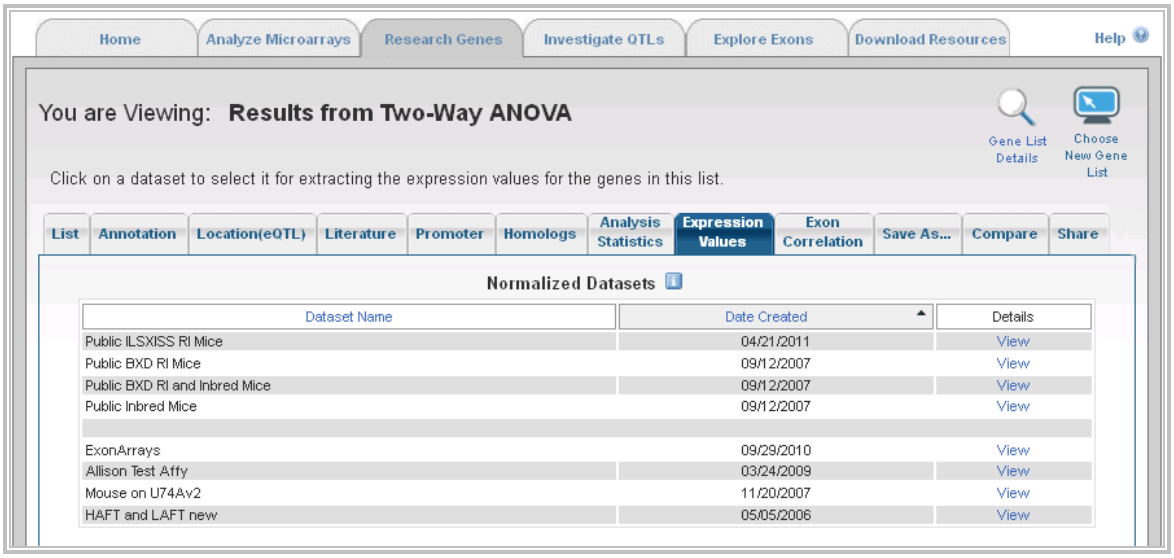
Analyze Microarrays tab:

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* tab displays.
2. Click **View expression values for a list of genes in a dataset** in the *What would you like to do* section. The *Expression Values* page displays.



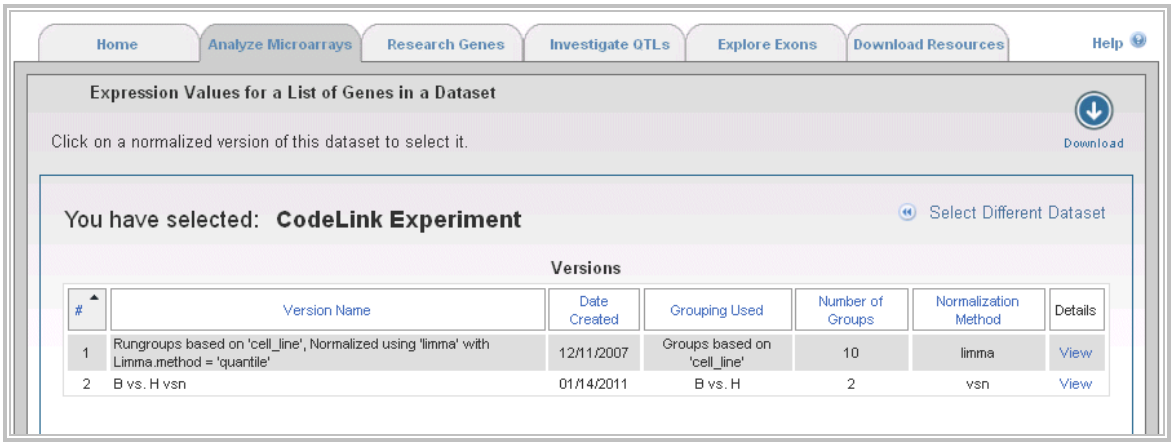
Research Genes tab:

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view gene expression data.
4. Click the **Expression Values** tab. The *Expression Values* page displays.

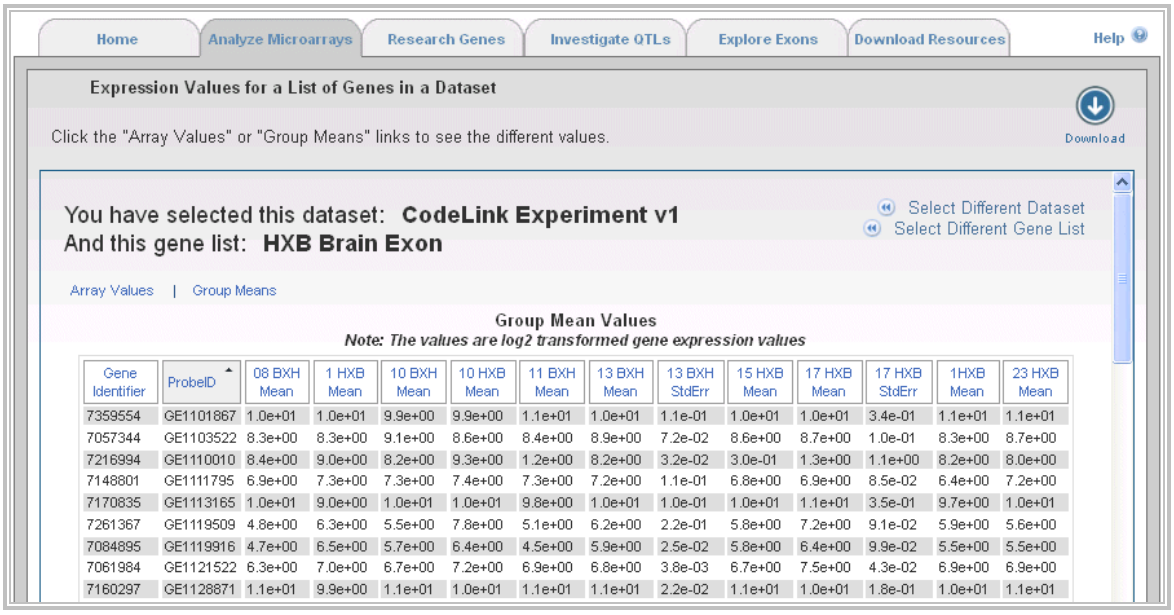


On the *Expression Values* page:

1. **OPTIONAL:** Click the **View** link in the **Details** column to view dataset details such as name, description, organism, arrays in dataset, and more. See "Viewing Dataset Details" for more information.
2. Click the **dataset** for which you want to see gene expression data. A page displays the normalized versions of that dataset.



3. Click the **normalized version** for which you want to see gene expression data. A page displays gene lists for that version.
4. Click the **gene list** for which you want to see gene expression data. The *View Gene Expression* page displays.



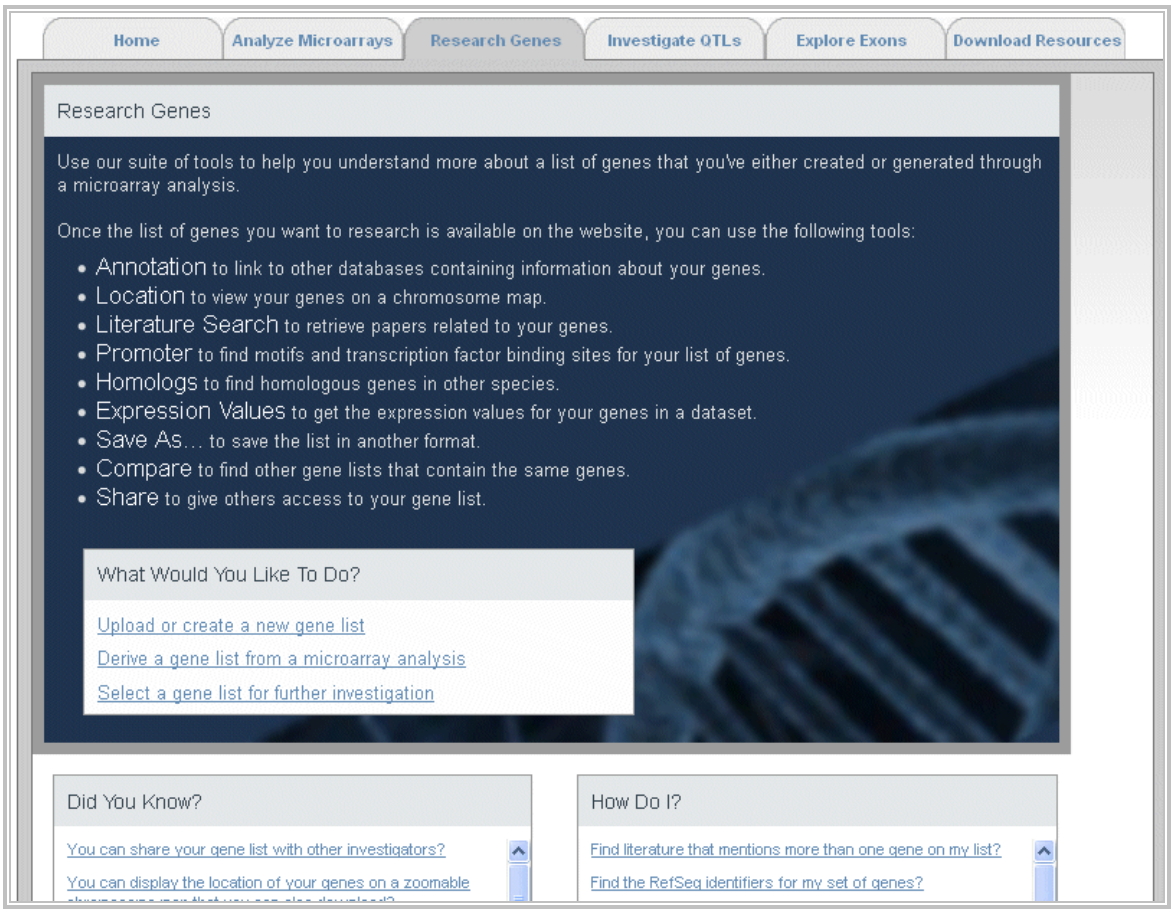
Gene Identifier	ProbeID	08 BXH Mean	1 HXB Mean	10 BXH Mean	10 HXB Mean	11 BXH Mean	13 BXH Mean	13 BXH StdErr	15 HXB Mean	17 HXB Mean	17 HXB StdErr	1HXB Mean	23 HXB Mean
7359554	GE1101867	1.0e+01	1.0e+01	9.9e+00	9.9e+00	1.1e+01	1.0e+01	1.1e-01	1.0e+01	1.0e+01	3.4e-01	1.1e+01	1.1e+01
7057344	GE1103522	8.3e+00	8.3e+00	9.1e+00	8.6e+00	8.4e+00	8.9e+00	7.2e-02	8.6e+00	8.7e+00	1.0e-01	8.3e+00	8.7e+00
7216994	GE1110010	8.4e+00	9.0e+00	8.2e+00	9.3e+00	1.2e+00	8.2e+00	3.2e-02	3.0e-01	1.3e+00	1.1e+00	8.2e+00	8.0e+00
7148801	GE1111795	6.9e+00	7.3e+00	7.3e+00	7.4e+00	7.3e+00	7.2e+00	1.1e-01	6.8e+00	6.9e+00	8.5e-02	6.4e+00	7.2e+00
7170835	GE1113165	1.0e+01	9.0e+00	1.0e+01	1.0e+01	9.8e+00	1.0e+01	1.0e-01	1.0e+01	1.1e+01	3.5e-01	9.7e+00	1.0e+01
7261367	GE1119509	4.8e+00	6.3e+00	5.5e+00	7.8e+00	5.1e+00	6.2e+00	2.2e-01	5.8e+00	7.2e+00	9.1e-02	5.9e+00	5.6e+00
7084895	GE1119916	4.7e+00	6.5e+00	5.7e+00	6.4e+00	4.5e+00	5.9e+00	2.5e-02	5.8e+00	6.4e+00	9.9e-02	5.5e+00	5.5e+00
7061984	GE1121522	6.3e+00	7.0e+00	6.7e+00	7.2e+00	6.9e+00	6.8e+00	3.8e-03	6.7e+00	7.5e+00	4.3e-02	6.9e+00	6.9e+00
7160297	GE1128871	1.1e+01	9.9e+00	1.1e+01	1.0e+01	1.1e+01	1.1e+01	2.2e-02	1.1e+01	1.0e+01	1.8e-01	1.0e+01	1.1e+01

5. You can:
    - Click the **Array Values** link to view the individual array values.
    - Click the **Group Means** link to view the group mean values (this is the default view).
    - Click **Download** to save the group means as well as the individual array values. Follow the instructions that display as you open or save the files. These instructions vary depending on your Internet browser (e.g., Internet Explorer, Firefox, Safari).

Researching Genes

The **Research Genes** tab allows you to create, delete, upload, and download gene lists, as well as access tools for interpreting them. Choose an option to get started:

- Upload or type in a new gene list. See "Creating a Gene List Overview" for details.
- Create a gene list from a micorarray analysis. See "Analyzing Datasets" for details.
- Select a gene list for further investigation.See "Viewing Gene Lists"for details.



Gene list data security requirements depend on the origin of the gene lists. The owner of a gene list can give other users permission to see gene lists that they own. If you have permission to view a gene list, because you own it or otherwise, you can also download that gene list.

Viewing Gene Lists

You can view and work with gene lists.

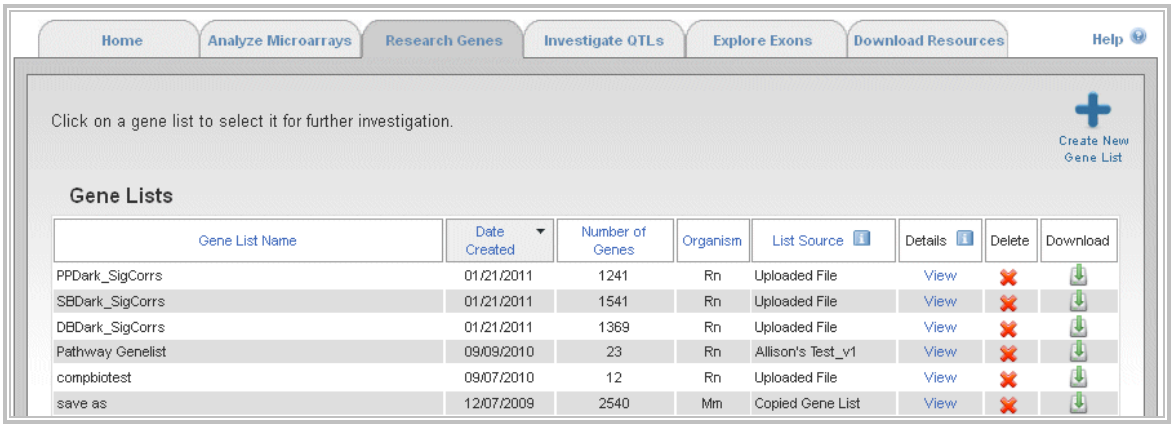
1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.

   Note: You can only view gene lists that you have created or to which you have been granted access. See "Sharing a Gene List".

You can:

- Click the **Create New Gene List** button to upload or manually create a new gene list.

- Click the **View** link in the **Details** column to view gene list details. See "Viewing Gene List Details" for more information.
- Click the **Delete** icon to delete a gene list.
- Click the **Download** icon to download a gene list.
- Click on a row to view the gene list.



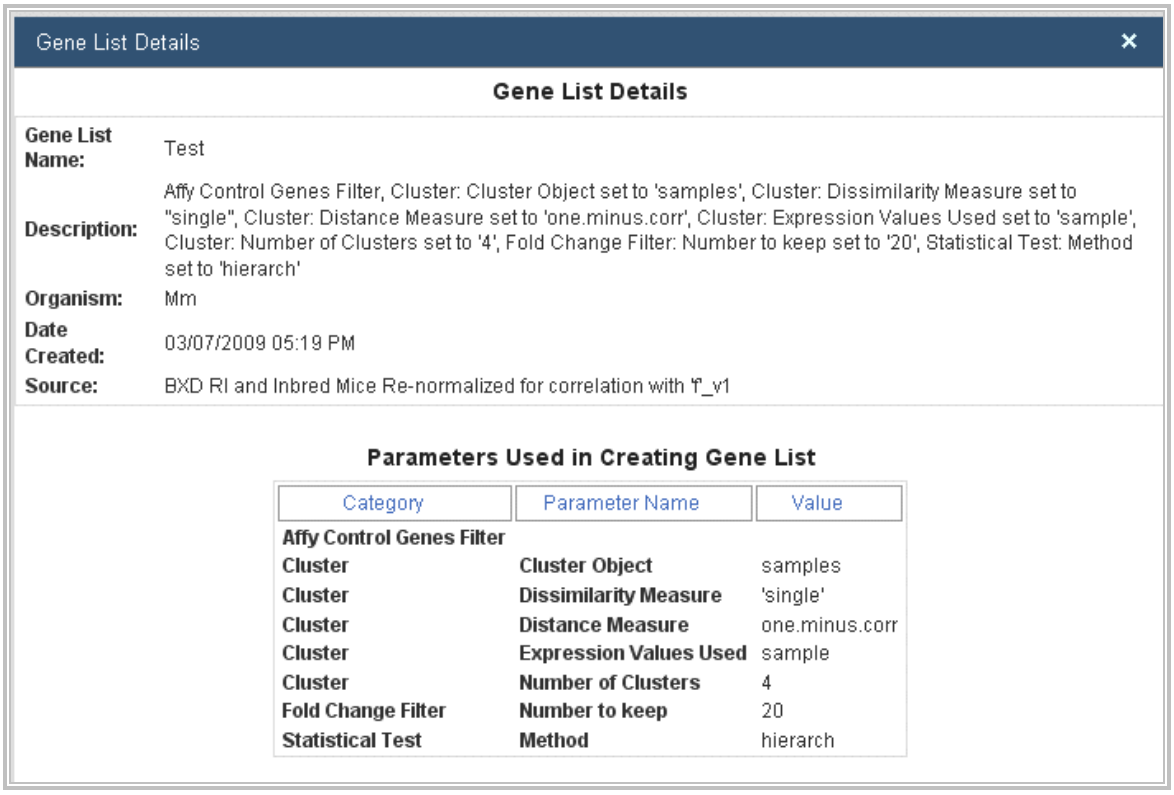
When you click a gene list, the following tabs display details for that gene list:

- **List** - Shows the list of genes in the selected gene list.
- **Annotation**- Allows you to perform annotation on a gene list.
- **Location** - Allows you to view the physical location and transcriptional control locations (eQTL) of your genes on an interactive chromosome map.
- **Literature** - Allows you to perform a literature search for a gene list
- **Promoter Analysis** - Allows you to run promoter analysis using oPOSSUM, MEME or Upstream Sequence Extraction on a gene list.
- **Homologs** - Allows you to obtain information regarding chromosomal location in other genomes.
- **Pathways** - Allows you to analyze the hypothetical impact of differences in transcription levels of a set of genes on signaling pathways defined by KEGG (Ogata et al 1999). The **Pathways** tab only displays when a full change or correlation coefficient is available.
- **Analysis Statistics** - Allows you to view the raw p-values and the adjusted p-values from the statistical analysis performed to generate this gene list. Depending on the number of groups and the type of analysis used, it may also display group means, F-statistic, mean intensity, correlation coefficient, difference in log base 2 intensity, t-statistic, or parameter estimates.
- **Expression Values** - Allows you to view normalized expression values of your genes in any data set.
- **Exon Correlations** - Allows you to an create an exon correlation heatmap for a specific gene, species, and tissue.
- **Save As** - Allows you to save gene lists translated to other types of identifiers.
- **Compare** - Allows you to compare gene lists, with the option to create, for example, unions or intersections of the gene identifiers.
- **Share** - Allows you to view users who have access to a gene list and to give permission to other users to access your gene list.

Viewing Gene List Details

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the **View** link in the **Details** column beside a gene list to view the details. You can also view the gene list details when you click the magnifying glass that displays after you select the gene list.

The resulting *Gene List Details* page provides basic information about a gene list such as name, description, organism, date created, and source. It also shows the filters, normalization parameters, and statistical analysis that were used to create the gene list.



Creating a Gene List Overview

You can create a gene list from:

- A differential expression analysis of a dataset.
- A correlation analysis of a dataset.
- A cluster analysis of a dataset.

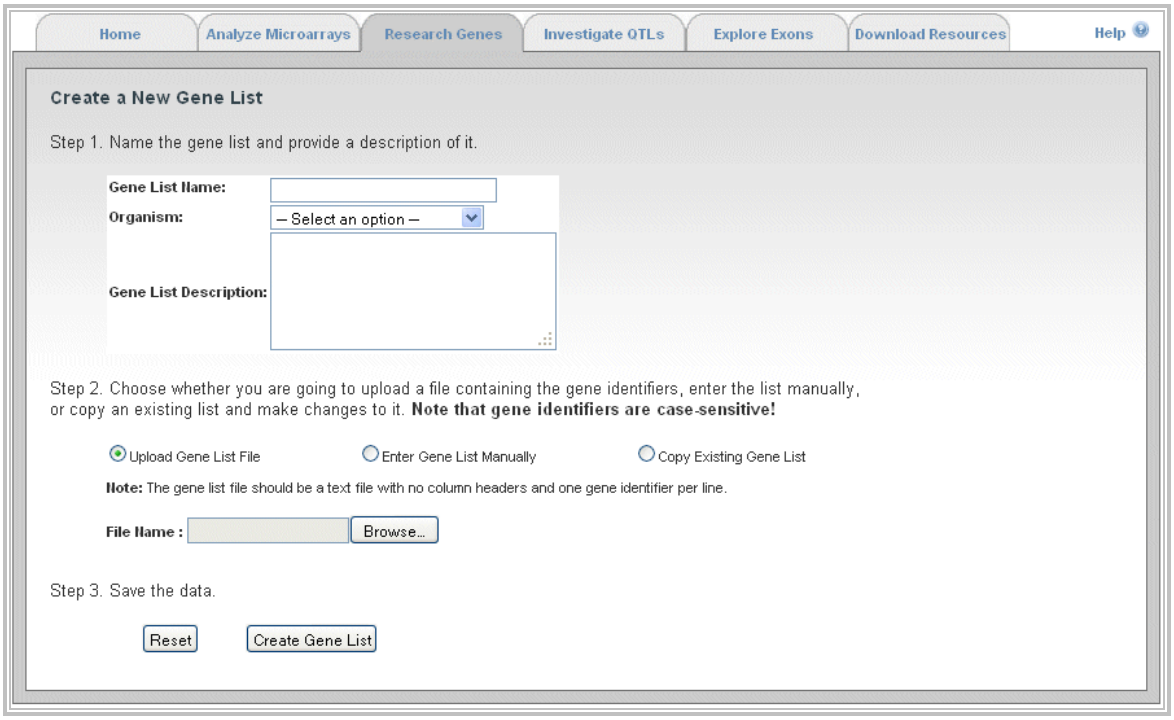
You can also manually enter, copy, or upload gene lists. See:

- "Manually Entering a Gene List" on page 97
- "Uploading a Gene List" on page 131
- "Copying a Gene List" on page 99

Uploading a Gene List

You can upload an existing gene list to the PhenoGen website for analysis. Your gene list must be a text file and must contain only one gene identifier per line or it will not upload correctly. When you upload gene list data, you can give other users permission to see the gene list.

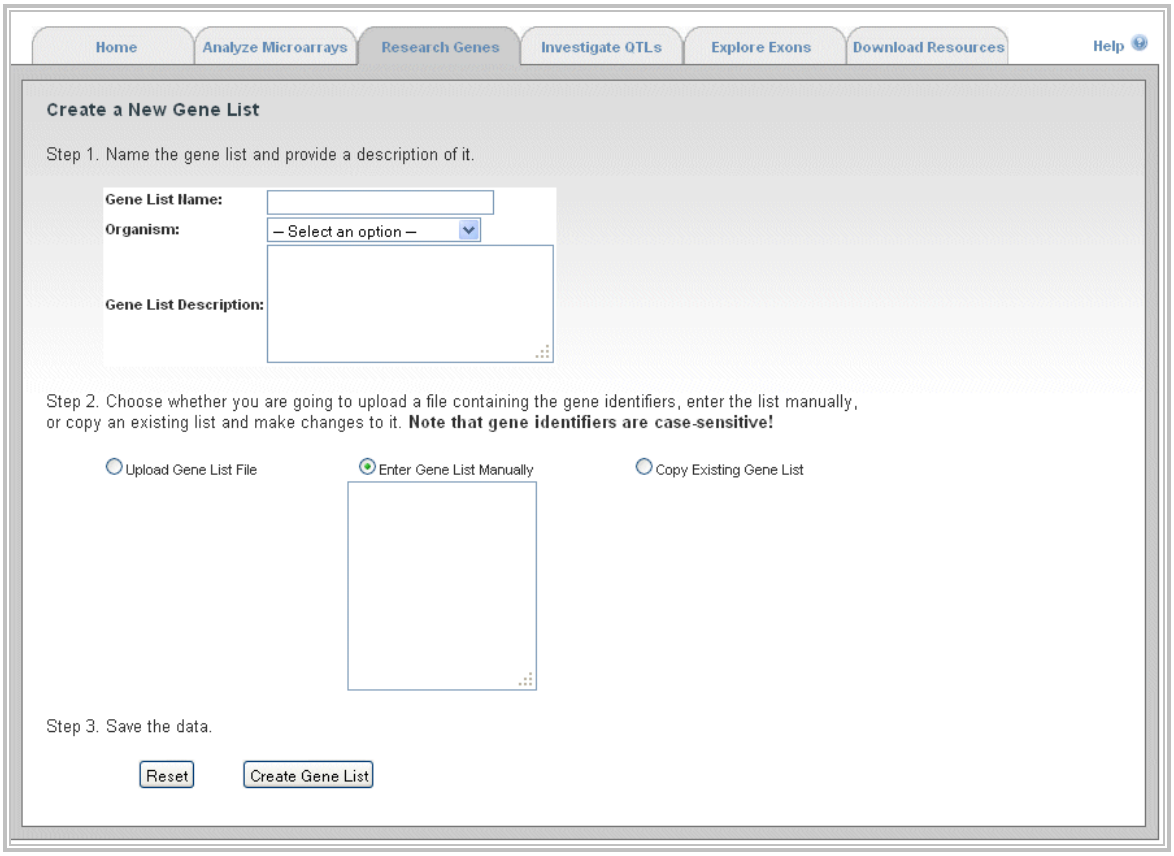
1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Upload or create a new gene list** in the *What would you like to do* section. The *Create a New Gene List* page displays.



3. Enter the **Gene List Name**.
4. Select the **Organism** from the drop-down list.
5. Enter the **Gene List Description**.
6. Make sure **Upload Gene List File** is selected.
7. Click **Browse**, and follow the instructions to select the gene list you want to upload.
8. Click **Create Gene List**. The new gene list is created.

Manually Entering a Gene List

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Upload or create a new gene list** in the *What would you like to do* section. The *Create a New Gene List* page displays.



3. Enter the **Gene List Name**.
4. Select the **Organism** from the drop-down list.
5. Enter the **Gene List Description**.
6. Make sure **Enter Gene List Manually** is selected.
7. Type in the genes to include in the gene list. You must separate the genes with a space or by pressing **Enter** on the keyboard, or they cannot be processed correctly. Genes can be entered in the following formats:

Gene IDs	Example	Description
Affymetrix 3' ID	1416283_at	
Affymetrix Exon Probe-set ID	420693	
Affymetrix Exon Tran-script ID	781497	
CodeLink ID	NM_009775_Probe1 GE34729	

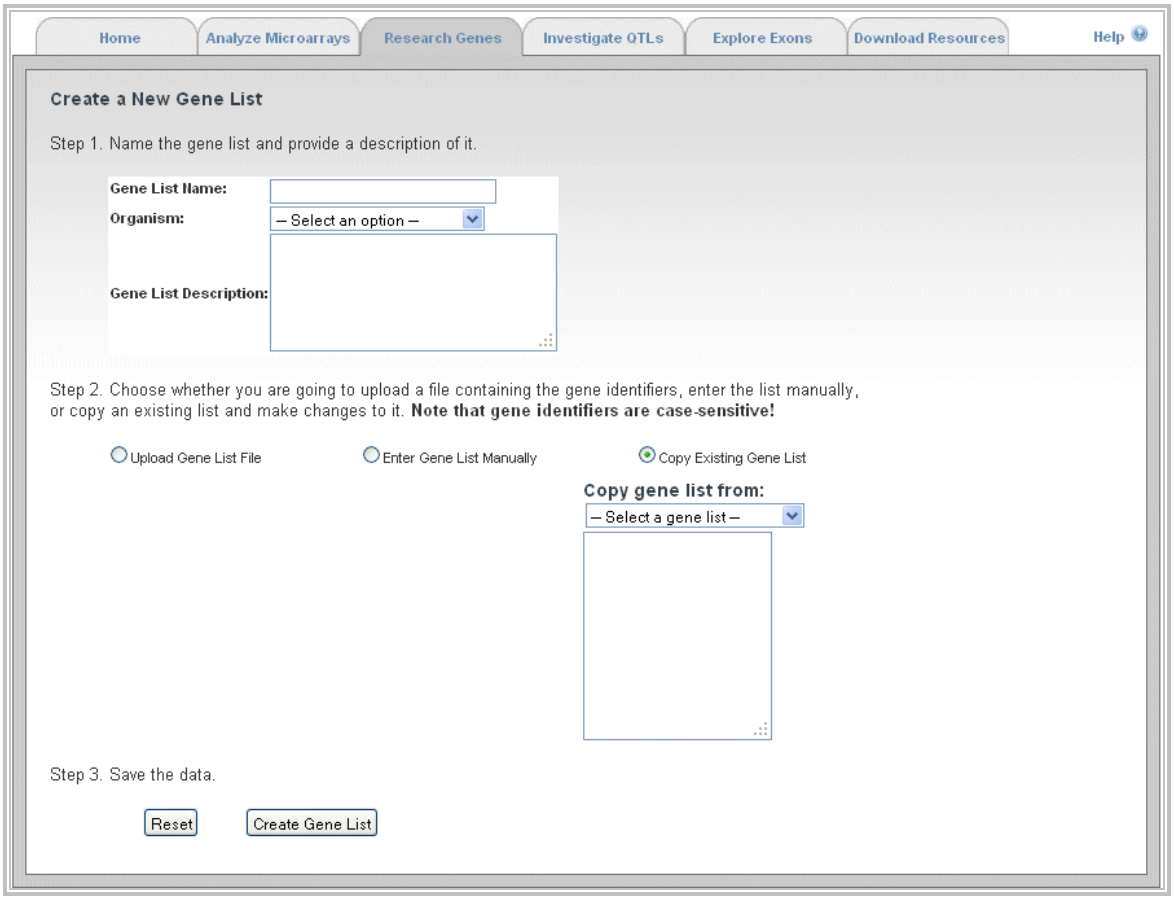
Ensembl ID	ENS-MUSG00000022962	Begins with ENS, followed by the three-letter organism, followed by G, followed by a number.
Entrez Gene ID	14450	Contains three or more numbers.
FlyBase Gene ID	FBgn0013277	All IDs are preceded by FBgn and have numbers following.
Gene Symbols	Grin1	Official gene symbols.
MGI ID	MGI:95654	
NCBI/EMBL RNA accession number	AK146355	Begins with one or two letters, followed by multiple numbers.
RefSeq RNA accession number	NM_010256	Begins with two capital letters, followed by an underscore, followed by numbers.
RGD ID	2203	Contains four or more numbers.
UniGene ID	Mm.4505	Contains an organism prefix (Dm, Hs, Rn, or Mm), followed by a period, followed by numbers.

Genes not entered in the formats above may not be recognized by the PhenoGen website.

8. Click **Create Gene List** when you are done. Your gene list displays on the *Research Genes* page.

Copying a Gene List

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Upload or create a new gene list** in the *What would you like to do* section. The *Create a New Gene List* page displays.



3. Enter the **Gene List Name**.
4. Select the **Organism** from the drop-down list.
5. Enter the **Gene List Description**.
6. Make sure **Copy Existing Gene List** is selected.
7. Select a gene list from the drop-down list. The gene list populates the box.
8. Type in new genes and delete existing genes as necessary.
9. Click **Create Gene List**. The new gene list is created.

Annotating Gene Lists

Gene lists on the PhenoGen website can contain identifiers from any source and can be translated using a tool called iDecoder. This tool translates identifiers to and from the following identifier types:

- AFFYMETRIX 3' PROBESET ID
- AFFYMETRIX EXON PROBESET ID
- AFFYMETRIX EXON TRANSCRIPT ID
- CODELINK PROBE ID
- ENSEMBL ID
- ENTREZ GENE ID
- FLYBASE ID
- GENE SYMBOL
- MGI ID
- REFSEQ PROTEIN ID
- REFSEQ GENE ID
- RGD ID
- SWISSPROT ID
- UNIGENE ID

The data used to translate these values comes from information downloaded from the following organizations:

- Affymetrix
- CodeLink (GE Healthcare)
- Ensembl
- FlyBase
- MGI
- NCBI
- RGD
- Swissprot

Your uploaded gene lists can be a mixture of many ID types, but all identifiers in a gene list must be for the same organism. For example, you can upload a gene list that contains an Affymetrix probe set ID, an official gene symbol, an Entrez Gene ID, and a RefSeq Gene ID, and iDecoder translates them into the identifier types appropriate for the selected tool.

iDecoder is the underlying program for the annotation tools on the PhenoGen website that maps gene identifiers between databases and . For instance, if database 1 contains entry A and database 2 contains entry B, and both A and B refer to entry C in database 3, but not to each other, iDecoder identifies that A and B are related. The method is very efficient in unearthing previously unknown equivalent IDs.

There are two levels of annotation available on the PhenoGen website:

- "Basic Annotation" on page 100
- "More Annotation" on page 101

Basic Annotation

Basic annotation displays links to the most popular databases for each of the identifiers in a gene list. In addition to general annotation, the Basic Annotation tool also provides information on expression QTL (eQTL), based on mouse or rat data. A **QTL** column displays in the Basic Annotation table and a **PhenoGen eQTL** link displays when the gene list entry matches either a probe ID or gene symbol in the eQTL data.

Expression QTL (eQTL)

The purpose of expression QTLs is to determine the location in the genome that controls the transcription level of a gene. eQTLs are calculated using traditional QTL techniques where the quantitative trait of interest is the expression level of a gene as measured by microarray analysis. On the PhenoGen website, eQTLs have been calculated for both mouse and rat data. When you run Basic Annotation on gene lists for either of these species, the eQTLs are reported. See "Expression QTL Derivation".

In the expression QTL table, the physical location of the probe set ID is shown, along with the location of the marker that represents the maximum LOD score for that transcript. The location of the marker with the maximum LOD score indicates the region of transcriptional control. For this analysis, if the physical location of a gene is near the location of transcriptional control, the gene is considered to be cis (locally)-regulated. Otherwise, the gene is considered trans (distally)-regulated. The physical location of probe(set)s were obtained using the BLAT software (http://www.kentinformatics.com) and mapping on the NCBI m37 mouse genome assembly and the RGSCv3.4 genome assembly obtained from the UCSC Genome Browser (http://-/genome.ucsc.edu).

Allen Brain Atlas

You can obtain the regional expression pattern of a given gene in the brain of a C57BL/6J mouse by clicking on the link provided in the **Allen Brain Atlas** column in the Basic Annotation table. These links are available ONLY for mouse gene lists. The Allen Brain Atlas (Lein et al., 2007) is an open-access database of gene expression in the C57BL/6J brain tissue. This database was created by the Allen Institute for Brain Science (Seattle, Washington) and contains data for genome-wide RNA expression obtained using high-throughput, *in situ* hybridization. In addition to the expression data, the Atlas also has a number of tools available for analyzing and visualizing the *in situ* images. Click the **Instructions** link, at the top of the Allen Brain Atlas column, to see basic instructions for viewing images on the Allen Brain Atlas. Comprehensive help documentation is available online at: http://www.brain-map.org.

References

1. Allen Brain Atlas [Internet]. Seattle (WA): Allen Institute for Brain Science. © 2004-2007. Available from: http://www.brain-map.org .
2. Lein ES et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature 445:168-176, doi:10.1038/nature05453.
3. Markram H (2007). Industrializing neuroscience. Nature 445:160-161.

More Annotation

More annotation options allow you to select a customized set of databases for the annotation of a given gene list. After you select a gene list, you can perform annotation and select one or more of the different databases to obtain further annotation. You can download this information. See "Using More Annotation Options" for details.

Performing Annotation

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to perform annotation.
4. Click the **Annotation** tab. The *Annotation* page displays and lists the equivalent ID from five different ID types:
    - Official Symbol
    - RefSeq
    - MGI (Mouse Genome Informatics), RGD (Rat Genome Database), or Fly Base

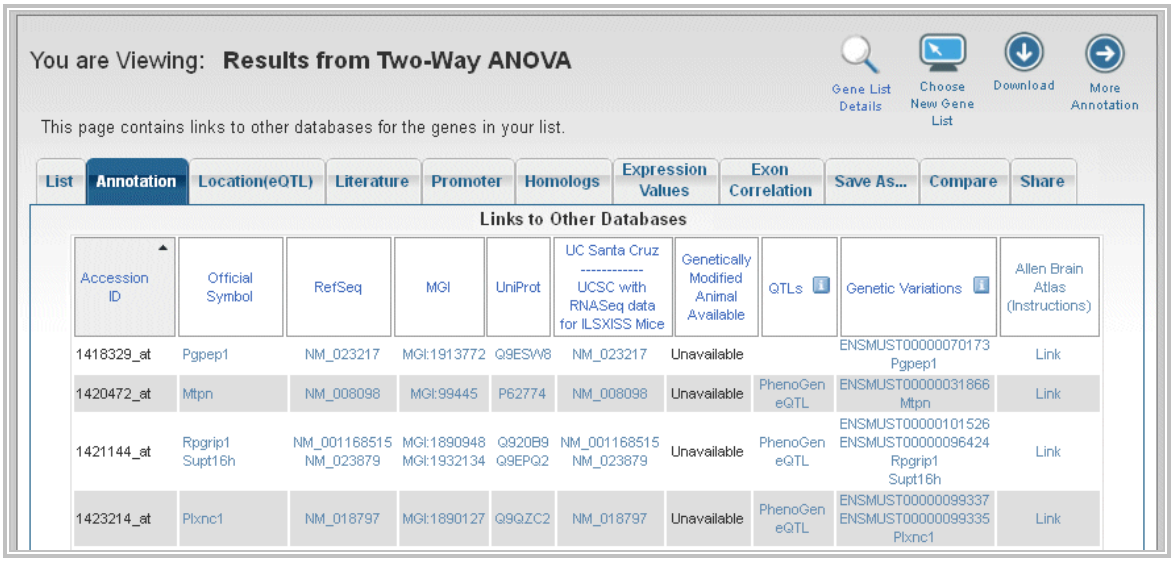
- UniProt (Swiss Prot)
- UC Santa Cruz (University of California, Santa Cruz)

It also lists:

- Genetically Modified Animal Available
- QTLs
- Genetic Variations
- Allen Brain Atlas (if applicable)

You can:

- Click the **Download** icon to download annotation information for the gene list.
- Click any link in a cell to open the website or information page for the ID you selected. If there are multiple links in a cell, each link opens a related page.



Accession ID	Official Symbol	RefSeq	MGI	UniProt	UC Santa Cruz ---------- UCSC with RNASeq data for ILSXISS Mice	Genetically Modified Animal Available	QTLs	Genetic Variations	Allen Brain Atlas (Instructions)
1418329_at	Pgpep1	NM_023217	MGI:1913772	Q9ESW8	NM_023217	Unavailable		ENSMUST00000070173 Pgpep1	Link
1420472_at	Mtpn	NM_008098	MGI:99445	P62774	NM_008098	Unavailable	PhenoGen eQTL	ENSMUST00000031866 Mtpn	Link
1421144_at	Rpgrip1 Supt16h	NM_001168515 NM_023879	MGI:1890948 MGI:1932134	Q920B9 Q9EPQ2	NM_001168515 NM_023879	Unavailable	PhenoGen eQTL	ENSMUST00000101526 ENSMUST00000096424 Rpgrip1 Supt16h	Link
1423214_at	Plxnc1	NM_018797	MGI:1890127	Q9QZC2	NM_018797	Unavailable	PhenoGen eQTL	ENSMUST00000099337 ENSMUST00000099335 Plxnc1	Link

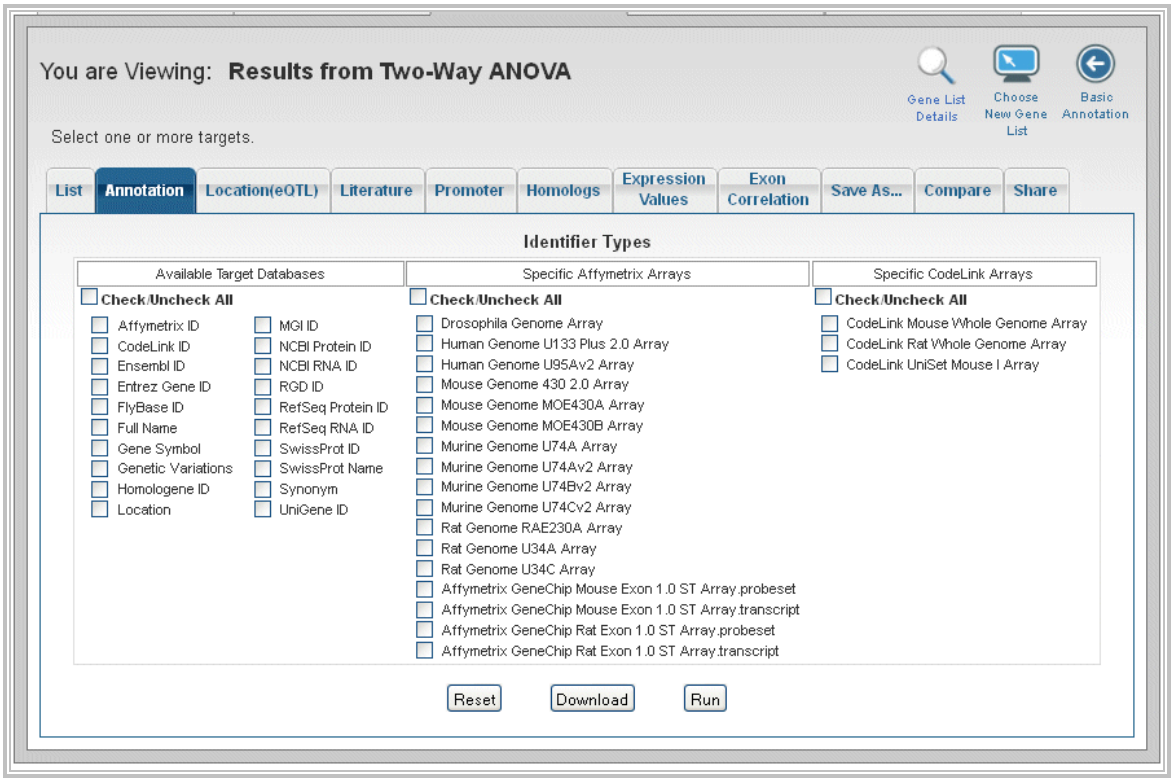
The Accession IDs display on the left, and the equivalent IDs for these genes from six specified resources are returned; Official gene symbol, NCBI's RefSeq, Jackson Laboratory's Mouse Genome Informatics (MGI), SwissProt's UniProt, University of California Santa Cruz, and Ensembl's TranscriptSNPView. These corresponding IDs are linked to their respective databases. When you click a link, further annotation is displayed in a new window. In addition to these links, if there are any genetically modified mice available for a particular gene, information from the database of genetically modified mice (maintained by MGI) is provided. Similarly, a link to PhenoGen eQTL data for a particular gene is also provided, if available. For mouse gene lists, a link to the Allen Brain Atlas may also be displayed for a specific gene. You can download the Basic Annotation information.

Using More Annotation Options

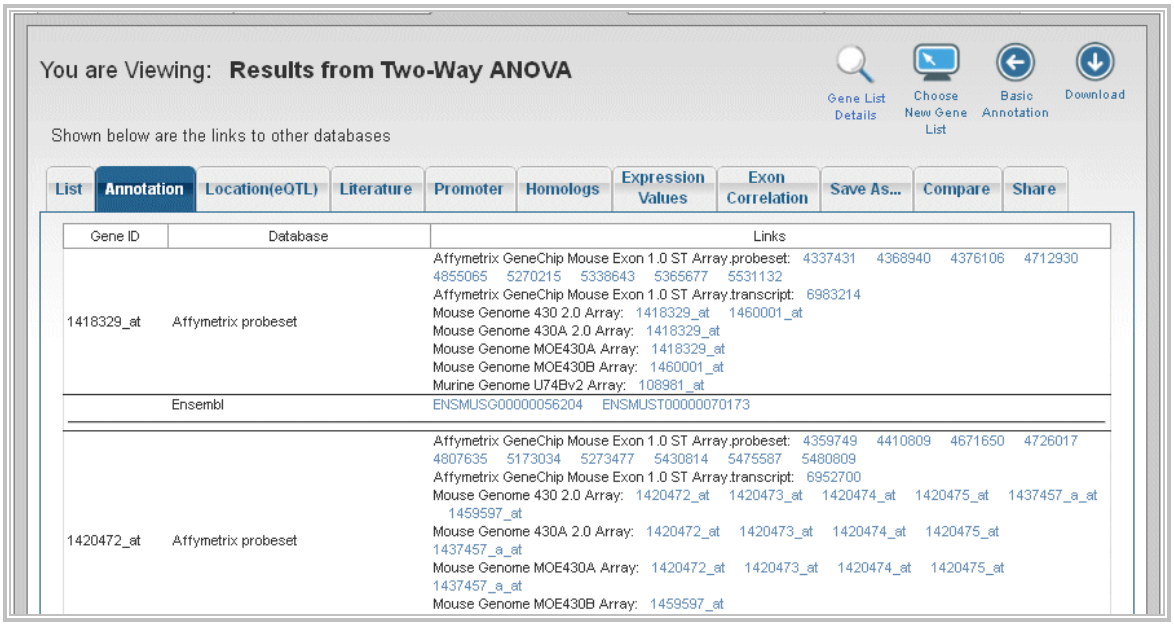
More annotation options allow you to select the gene list you want to annotate, then select the links you want iDecoder to search for.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to perform more annotation.
4. Click the **Annotation** tab.

5. Click **More Annotation**.
6. Select one or more target database links and array name(s), if desired.



7. Click **Download** to download the selected database(s) .
8. Click **Run** to view annotation results, including the Gene ID, the database(s) searched, and the links found in each database for your selections.



9. Click any of the different links to open the website for the link you selected.

Viewing Location and eQTL

The location graphic allows you to investigate and use expression QTL (eQTL) datasets. By default, only the physical locations are shown on the graphic. You can click the "Show locations of transcription control in brain" check box to display the locations of transcription control. These green arrows represent the genomic position(s) that control transcription of the candidate genes in the gene list (expression QTLs, or eQTLs).You can also change the significance level for displaying eQTLs

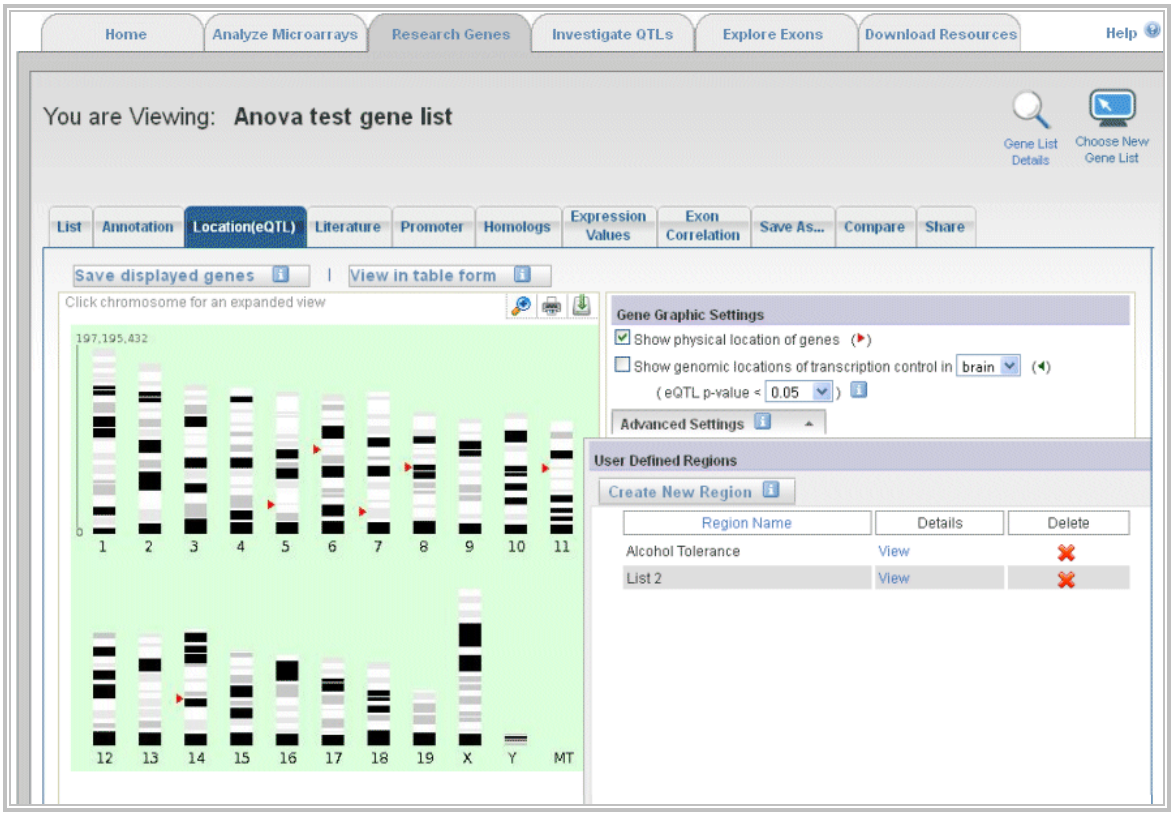
You can view the location and eQTL data for a gene list in two ways:

From the *Research Genes* tab:

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view location and eQTL.
4. Click the **Location** tab. The *Location* page displays.

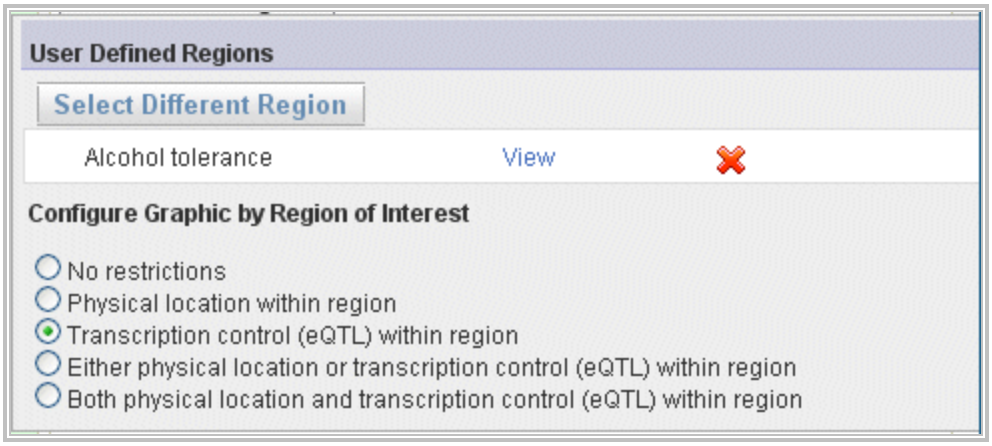
From the *Investigate QTL Regions* tab:

1. Click the **Investigate QTL Regions** tab. The *Investigate QTL Regions* page displays.
2. Click **View physical location and eQTL information about specific genes** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view location and eQTL. The *Location and eQTL* page displays.



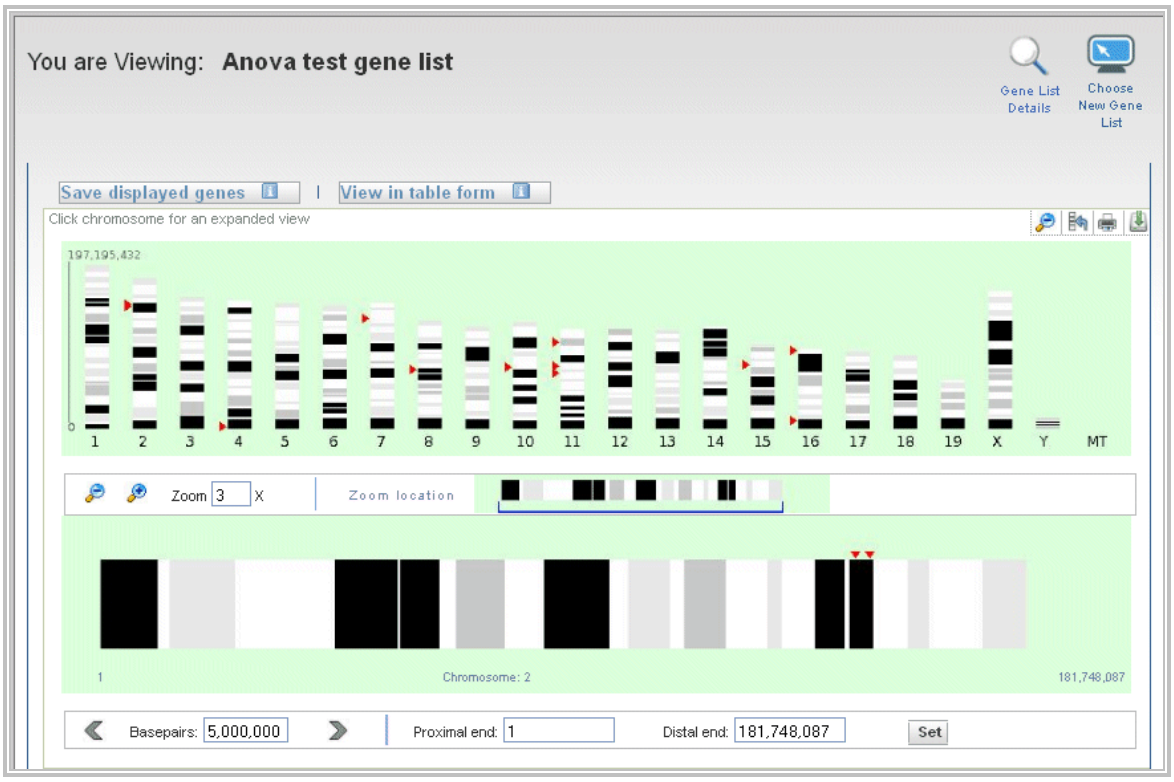
The data on the page looks the same, regardless of the way you access the page. You can:

- Click the **View in table form** link to see a table of all of the genes currently displayed on the graphic. If any of the probesets for a given gene have a significant eQTL, all of the probesets for that gene are displayed in the table, but only the most significant eQTL position ("Max LOD") and the associated marker locations are shown on the graphic.
    - Click the **View** link in the gene table to display individual LOD plots for each probeset. Copy these by right-clicking on the image.
    - Click the **Download** button to download the table into a tab-delimited text file.
    - Click the **Customize this view** button to change the default view of the table, which includes all probesets for a given gene.
        - Choose **Genes (and all associated probesets) from the [selected] list that meet the restriction criteria** to restrict the list to those with eQTLs (probesets) that overlap the selected bQTL intervals.
        - Select **Probesets that are in the [selected] gene list** to restrict the list to probesets (as opposed to genes) whose expression values are significantly correlated with the phenotype. These probesets are also identified by the asterisk in the first column of the table.
        - Selelct **Probesets that have eQTL p-value < [selected value]** to restrict the list to only the probesets passing the desired threshold.
        - Choose **Genes (and all associated probesets) from the [selected] list with probesets that did not meet the restriction criteria or were not considered in eQTL** to restrict the list to those that were not considered in eQTL.
- Click the **Download** icon  to download the chromosome map as a JPEG graphic.
- Click the **Save displayed genes** link to save the displayed genes as a gene list.
- Click the **Magnify**  icon in the graphic to expand the graphic to the full page.
- Click a chromosome in the graphic to show an expanded view. See "Expanded Chromosomal View".
- Set the *Gene Graphic Settings*:
    - **Show physical location of genes** - Choose this option to mark the physical location of genes with a red arrow.
    - **Show genomic locations of transcription control in...**Choose this option to mark the transcription control locations with a green arrow, and choose a corresponding eQTL p-value. For mouse, only brain is available. For rat, options are brain, heart, liver, and brown adipose.
- Click **Advanced Settings** to enter user-defined regions, such as QTLs. You can also:
    - Click **View** to see details about the specific region.
    - Click the **Delete** icon beside the region you want to delete.



Expanded Chromosomal View

The expanded chromosomal view zooms in on the specific location you choose in the graphic, shows the zoom location, and allows you to set basepair start (left) and end (right) positions.



Literature Search Overview

The Literature Search option is an automated literature search that you can tailor to your area(s) of interest by selecting a set of query terms. Query terms can be further organized into categories. The automated literature search tool searches abstract text and article titles that contain the gene ID, plus one or more of the query terms. The results of the search are organized by the user-defined categories and by gene ID or name. The return page provides the title, abstract, and PubMed link for each of the documents. The gene identifiers and keywords from the search are also highlighted to help you sort, read, and work through what is likely to be a large amount of text.

Co-reference Analysis

An additional feature of the literature search on the PhenoGen website is that publications are flagged if more than one of the genes within the gene list are mentioned in a single article. This allows you to easily identify genes that have previously shown a documented relationship.

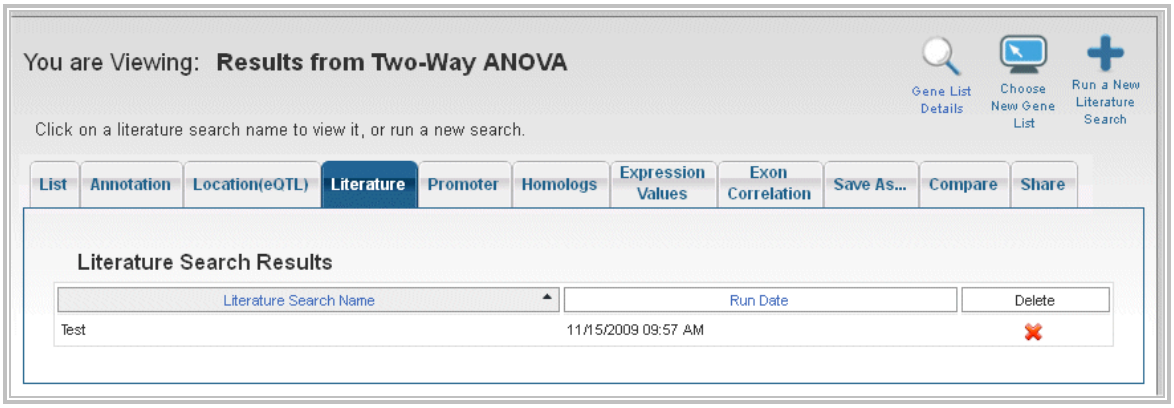
Performing a Literature Search

The Literature Search on the PhenoGen website allows you to perform a Literature Search for a gene list as well as the co-references for each gene in the list.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to perform a literature search
4. Click the **Literature** tab. The *Literature* page displays.

   You can:

   - Click the **Delete** icon to delete a literature search.
   - Click a literature search to view the results of the selected literature search. See "Viewing Literature Search Results".
   - Click **Run a New Literature Search** to begin a new search.



Creating a Literature Search

1. Click **Run a New Literature Search** on the Literature tab to begin a new search.
2. Select a value from the drop-down list in the **Categories** field. The category field provides you with a way to categorize your keywords.

   Note: The Category name is NOT used in the search, but is a required field. ONLY values in the keyword fields are used in the search.

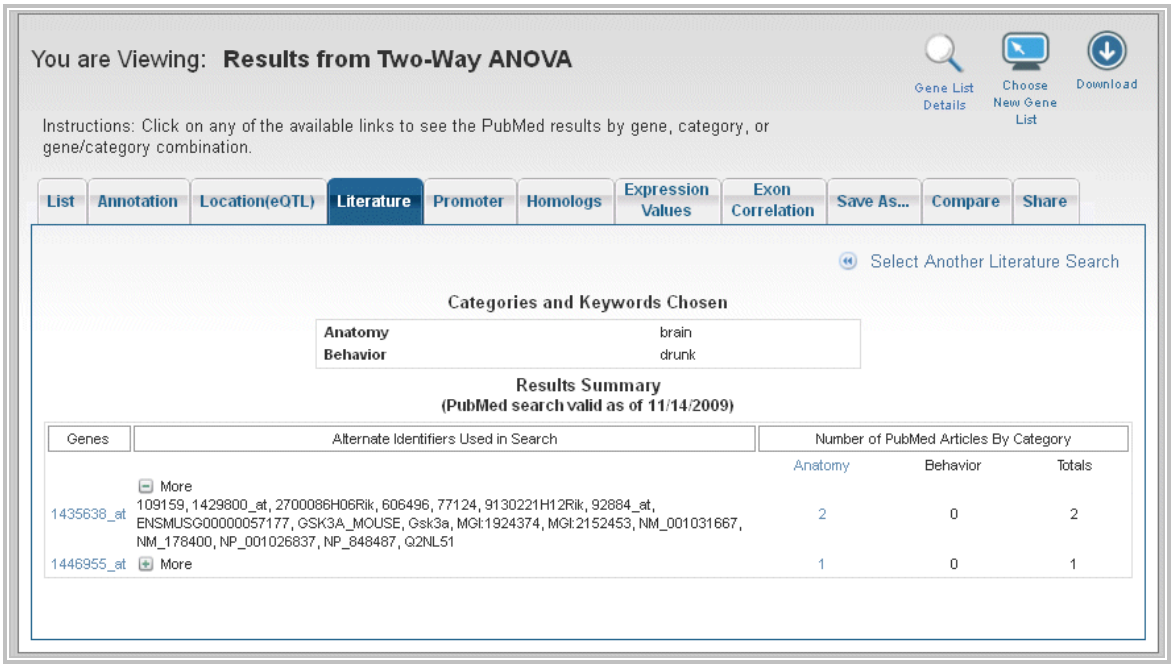
2. Enter a **keyword** or keywords for the category you selected.
3. **OPTIONAL:** Click **Clear Fields** to clear all the categories and keywords fields.
4. Enter a **Literature Search Name**.
5. Click **Submit Literature Search**. The Literature Search uses iDecoder to determine the synonyms for all genes in the gene list that is searched. The search looks at PubMed articles for all synonyms and keywords that were entered in the search page.

   Literature searches take time. When your results are available for viewing, an email is sent to the address you provided in the *Registration* page.

Viewing Literature Search Results

You can view literature search results for all literature searches you have performed. The results provide links to PubMed articles, organized by category. Links are also provided to articles where more than one gene is referenced in the same article. The PubMed results contain titles and the associated abstracts that are found for the search terms.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to see literature search results.
4. Click the **Literature** tab. The *Literature* page displays.
5. Click a literature search to view the results of the selected literature search.



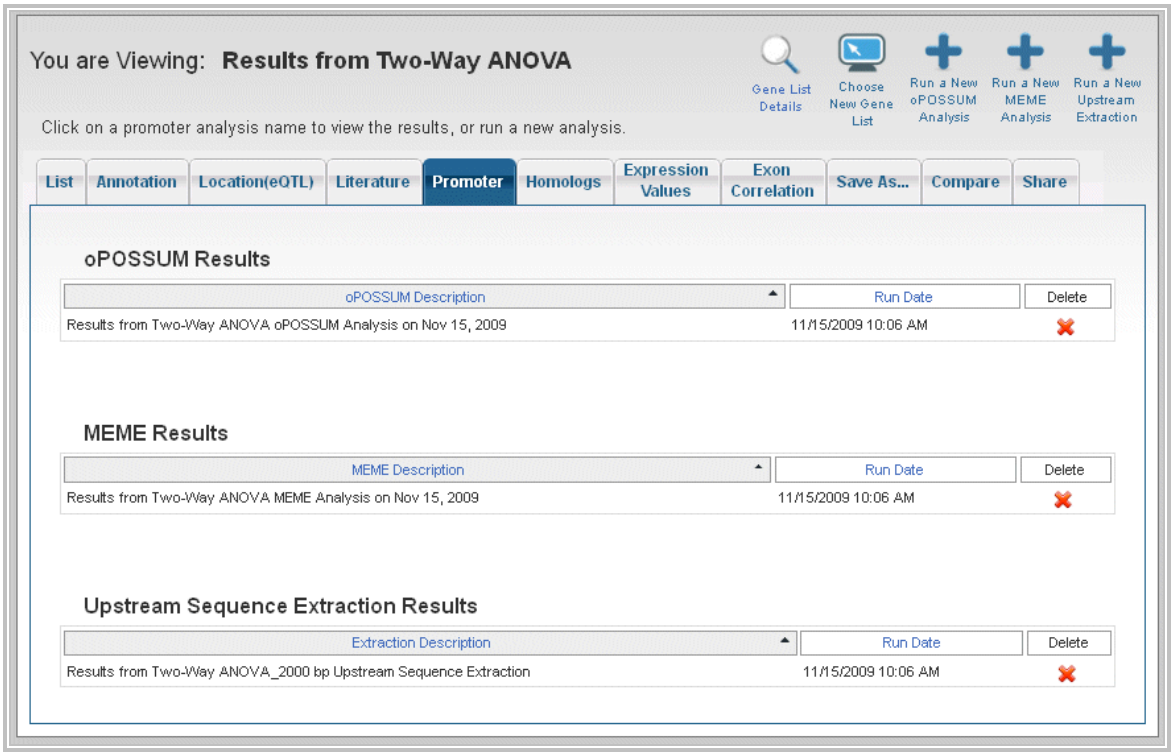
The *Literature Results* page shows the categories and keywords entered. It also displays the results summary and a list of coreferenced genes below the results summary. Note that gene names and keywords are highlighted when displaying the abstracts retrieved from a literature search. On the *Literature Results* page:

- Click any of the available **links** to see the PubMed results by gene, category, or gene/category combination.
- Click **More** to display the synonyms for a gene that were also listed in the search.
- Click the **Download** icon to download the results.

Promoter Analysis and Extraction

The PhenoGen website allows you to perform oPOSSUM and MEME promoter analysis, and upstream sequence extraction.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to perform a promoter analysis.
4. Click the **Promoter** tab. The *Promoter* page displays.



oPOSSUM Overview

is a tool for determining the over-representation of transcription factor binding sites (TFBS) within a set of (co-expressed) genes as compared with a pre-compiled (Ho Sui et al., 2005, Nucleic Acids Res 33(10):3154-64). The input is a set of gene identifiers and analysis . The system compares the number of hits for each selected TFBS on the target gene set against the background set. Two different measures of statistical significance are applied to determine which TFBS sites are over-represented in the target set. The of the analysis are displayed in a tabular form.

Notes:

- The PhenoGen website uses a customized version of oPOSSUM featuring a sub-set of input parameters.
- All matrices in the oPOSSUM database with a given minimum specificity are selected. These matrices are obtained from the JASPAR database.

Selection Criteria

Search Regional Level

This refers to the size of the region around the transcription start site (TSS) which was analyzed for TFBS sites. The background set was computed using a region extending a maximum of 5000 bp upstream and 5000 bp downstream of the TSS. During the background computation the upstream region was truncated to less than 5000 bp if it overlapped an upstream exon from another gene.

Conservation Level

To limit spurious TFBS sites, conservation with the aligned orthologous mouse sequence was used as a filter, and only sites which fell within these non-coding conserved regions were kept. A conserved region was defined as a span of some minimum length L within the human sequence which had a percent identity with the aligned mouse sequence of some minimum value X. The background set was pre-computed with three levels

of conservation filter. Level 1 corresponds to the top 10 percentile of non-coding conserved regions with an absolute minimum percent identity of 70%. Level 2 corresponds to the top 20 percentile with a minimum percent identity of 65% and level 3 corresponds to the top 30 percentile with a minimum percent identity of 60%.

Matrix Match Threshold

TFBS sites are scanned by sliding the corresponding position weight matrix (PWM) along the sequence and scoring it at each position. The threshold is the minimum relative score used to report the position as a putative binding site. The background set was computed using a threshold of 70%.

Statistical measure for over-representation

Two measures of statistical over-representation are available: a one-tailed Fisher exact probability and a Z-score.

One-tailed Fisher Exact Probability

The one-tailed Fisher exact probability compares the proportion of co-expressed genes containing a particular TFBS to the proportion of the background set that contains the site to determine the probability of a non-random association between the co-expressed gene set and the TFBS of interest. It is calculated using the hypergeometric probability distribution that describes sampling without replacement from a finite population consisting of two types of elements. Therefore, the number of times a TFBS occurs in the promoter of an individual gene is disregarded, and instead, the TFBS is considered as either present or absent.

Z-score

The Z-score uses a simple binomial distribution model to compare the rate of occurrence of a TFBS in the target set of genes to the expected rate estimated from the pre-computed background set.

For a given TFBS, let the random variable x denote the number of predicted binding site nucleotides in the conserved non-coding regions of the target gene set. Let B be the number of predicted binding site nucleotides in the conserved non-coding regions of the background gene set. Using a binomial model with n events, where n is the total number of nucleotides examined (i.e., the total number of nucleotides in the conserved non-coding regions) from the co-expressed genes, and N is the total number of nucleotides examined from the background genes, then the expected value of x is u = B * C, where C = n / N (i.e., C is the ratio of sample sizes). Then taking p = B / N as the probability of success, the standard deviation is given by s = sqrt(n * p * (1 - p)).

Let *x* be the observed number of binding site nucleotides in the conserved non-coding regions of the co-expressed genes. By applying the Central Limit Theorem and using the normal approximation to the binomial distribution with a continuity correction, the z-score is calculated as $z = (x - u - 0.5) / s$. Then, the probability of observing *x* or more binding site nucleotides in the conserved non-coding regions of the target genes, given the TFBS is not truly over-represented in the target genes, is the p-value associated with $Pr(Z >= z)$.

MEME Overview

The MEME (Multiple EM for Motif Extraction) search is based on occurrences of known motifs (transcription factor binding sites). There are many software options available to explore the occurrence of previously uncharacterized motifs. Although these have not been directly incorporated within the PhenoGen website as with oPOSSUM, they can easily be applied using other publicly available web servers.

A recent comprehensive review (Tompa et al., 2005, Nature Biotechnology 23:137) of such programs found that MEME (Bailey and Elkan, 1995, Proc. Int Conf Intell Syst Mol Biol 3:21), was one of the best performing algorithms on mouse data. Methods like MEME are optimal for analyzing sequences less than 2KB and it is not recommended to use longer lengths for such tools. Furthermore, many motif software webservers restrict the input data size. In addition to accessing MEME on the PhenoGen website, MEME can also be accessed at http://meme.sdsc.edu/meme/meme.html.

Upstream Sequence Extraction Overview

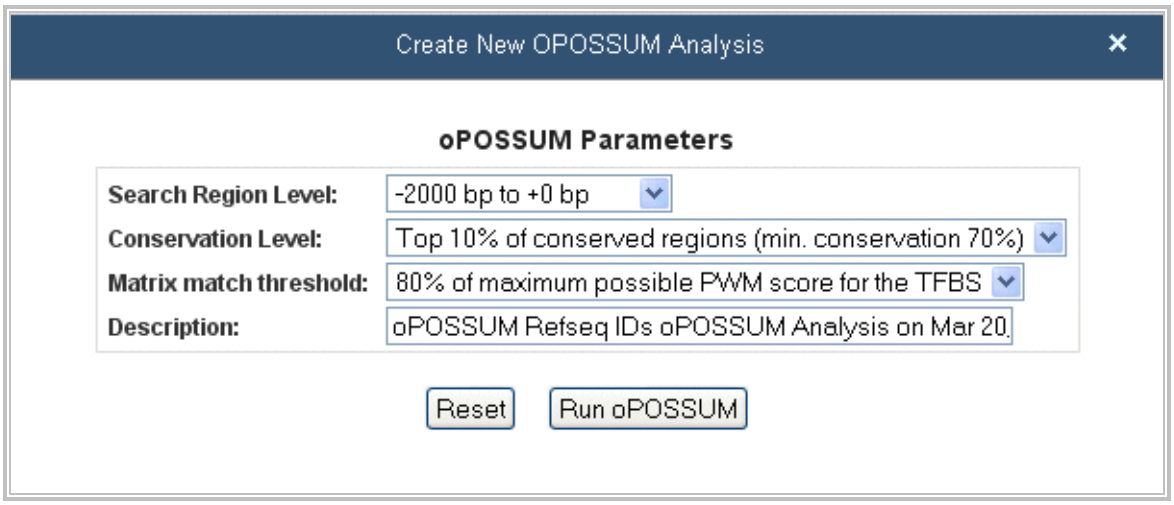
An important step in understanding the mechanisms that regulate the expression of genes is the ability to identify regulatory elements, i.e., the binding sites in DNA for transcription factors. Transcription factors are DNA binding proteins, typically upstream from, and close to, the transcription start site (TSS) of a gene, that modulate the expression of the gene by activating or repressing the transcription machinery.

Because there is a limited amount of information regarding the majority of the transcription factors and especially about their target binding sites (even in well-characterized organisms) you could focus on computational tools designed for the discovery of novel regulatory elements, where nothing is known a priori of the transcription factor or its preferred binding sites. If you provide a collection of sequences that correspond to the regulatory regions of genes that are believed to be co-regulated, the computational tool identifies short DNA sequence 'motifs' that are statistically over- or under-represented in these regulatory regions. Accurate identification of these motifs is very difficult because they are short signals (typically about 10 bp long) in the midst of a great amount of statistical noise (a typical input being one regulatory region of length 1,000 bp upstream of each gene). Also, there is marked sequence variability among the consensus binding sites of a given transcription factor, and the nature of the variability itself is not well understood.

There are numerous tools available for this task of motif prediction. They differ from each other mainly in their definition of what represents a motif and what would be an acceptable model for statistical over-representation of a motif. A comprehensive list of tools that could be used (table adapted from Tompa et al, 2005, Nature 23(1):137-144) is presented in "Supplementary Information" on page 154. This sequence information can be used to carry out TFBS analysis, off the PhenoGen website, using any of these tools. See "Promoter Analysis Tools".

Running oPOSSUM

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to run oPOSSUM.
4. Click the **Promoter** tab. The *Promoter* page displays.
5. Click **Run a new oPOSSUM analysis**. The *oPOSSUM* page displays.

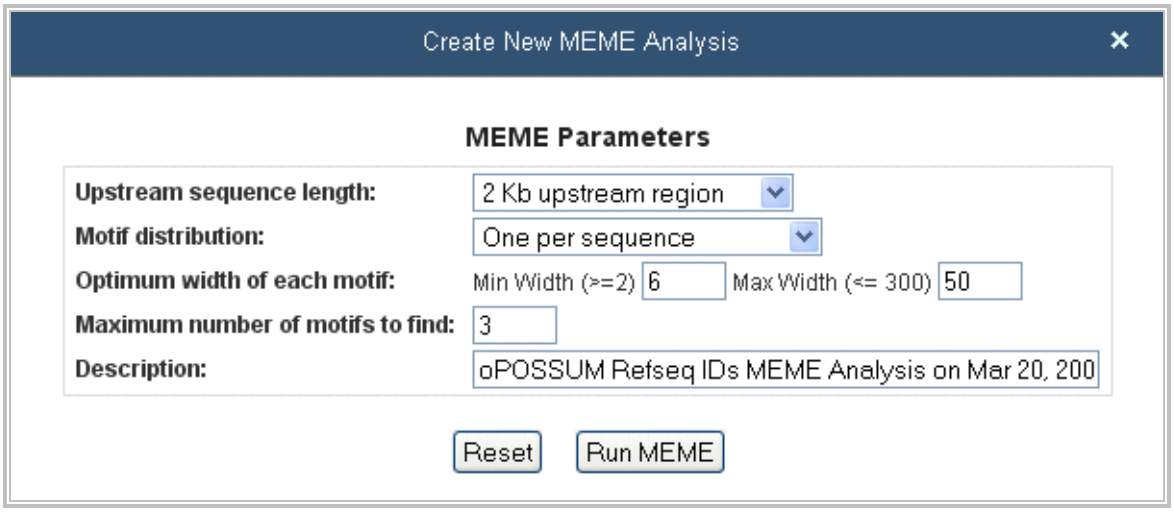


6. Change the parameters as necessary, using the drop-down lists.
   - Set the **Search Region Level**.
   - Set the **Conservation Level**.
   - Set the **Matrix Match Threshold**.

7. Change the **Description**, if appropriate.
8. Click **Run oPOSSUM**. Running oPOSSUM takes time. When your results are available for viewing, an email is sent to the address you provided in the *Registration* page.

Running MEME

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to run MEME.
4. Click the **Promoter** tab. The *Promoter* page displays.
5. Click **Run a new MEME analysis**. The *MEME* page displays.

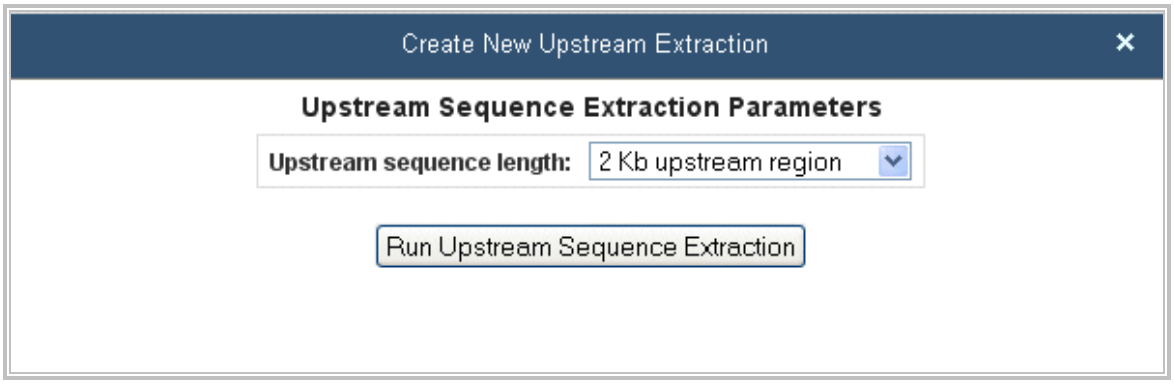


6. Select the **Upstream sequence length**.
7. Choose the **Motif distribution**.
8. Specify the **Minimum Width** and **Maximum Width** of each motif.
9. Specify the **Maximum number of motifs to find**.
10. Change the **Description**, if appropriate.
11. Click **Run MEME**. Running MEME takes time. When your results are available for viewing, an email is sent to the address you provided in the *Registration* page.

Running Upstream Sequence Extraction

The upstream sequence extraction tool is used to extract the upstream genome sequence of a particular gene.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to run Upstream Sequence Extraction.
4. Click the **Promoter** tab. The *Promoter* page displays.
5. Click **Run a new Upstream Extraction**. The *Upstream Extraction* page displays.

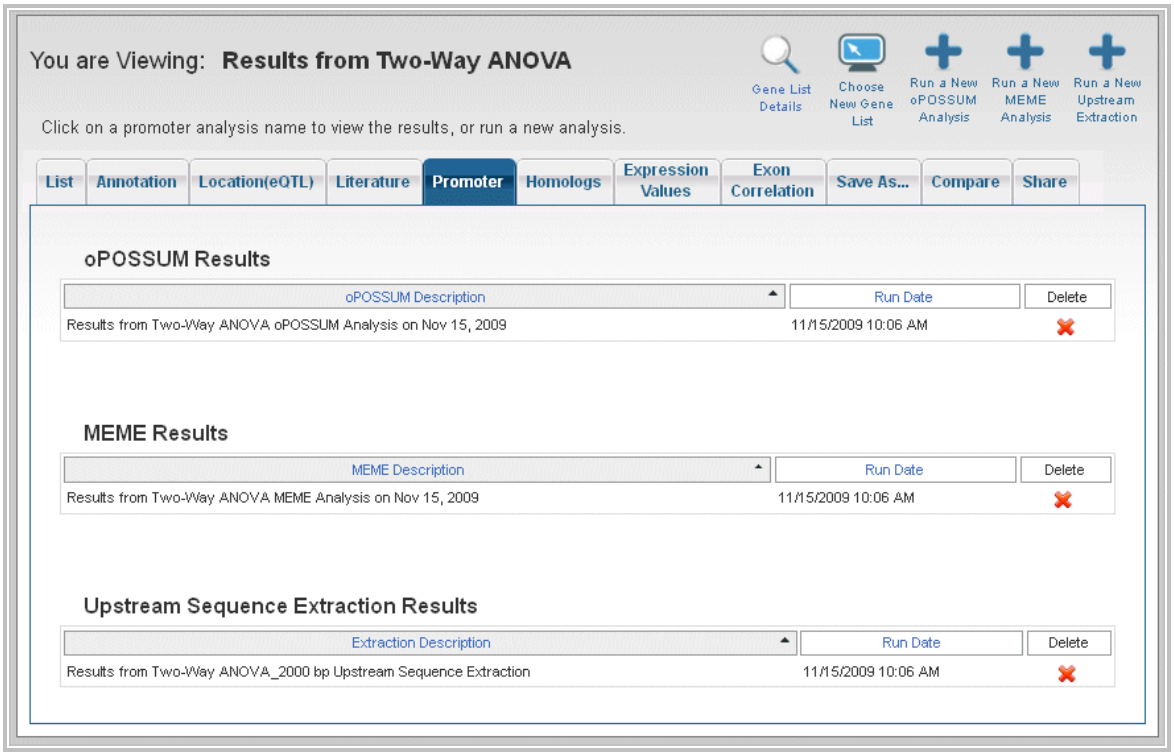


6. Select the **Upstream sequence length.**
7. Click **Run Upstream Sequence Extraction**. Running Upstream Extraction takes time. When your results are available for viewing, an email is sent to the address you provided in the *Registration* page.

Viewing Promoter Results

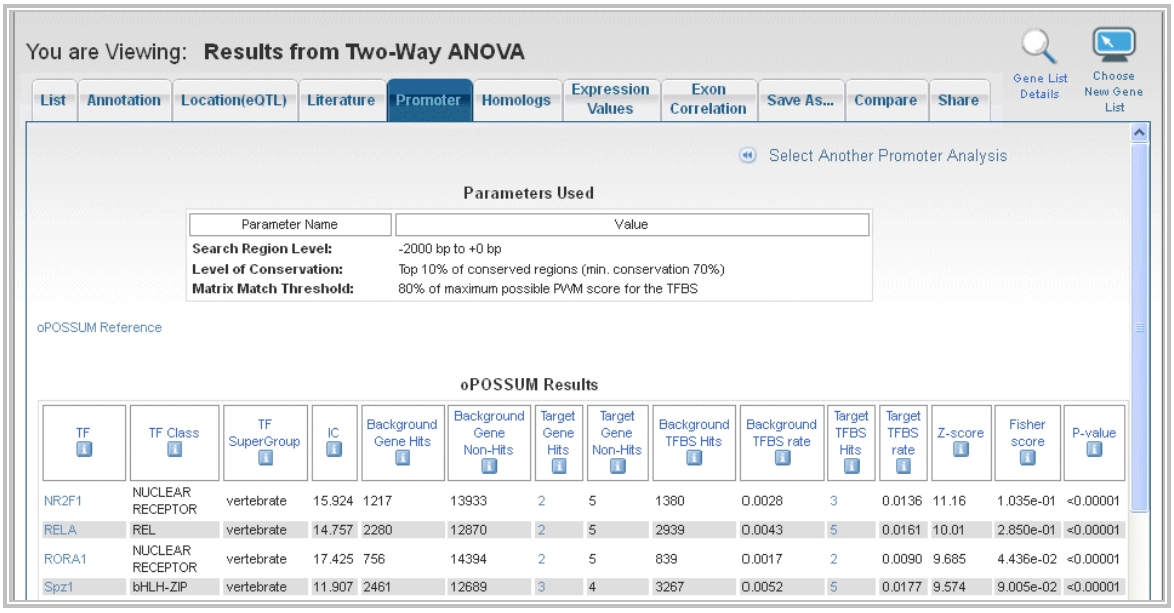
Three types of Promoter results may be available, depending on the Promoter analyses you chose to run: oPOSSUM, MEME, or Upstream Sequence Extraction.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view promoter results.
4. Click the **Promoter** tab. The *Promoter* page displays with results organized into tables by the type of analysis performed.
5. Do one of the following:
   - Click a promoter analysis to view the results of the selected analysis; oPOSSUM, MEME, or Upstream Sequence Extraction.
   - Click the **Delete** icon to delete the results of a promoter analysis.



Viewing oPOSSUM Results

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view oPOSSUM results.
4. Click the **Promoter** tab. The *Promoter* page displays with results organized into tables by the type of analysis performed.
5. Click a row in the *oPOSSUM Results* table. The *oPOSSUM Results* page displays the parameters used in the oPOSSUM analysis. It also displays transcription factors based on the One-Tailed Fisher Extract Probability Analysis and the Z-score Analysis.



oPOSSUM Results Table

The table contains the results from oPOSSUM, ordered by p-value from most to least significant (lower to higher p-value). The columns are:

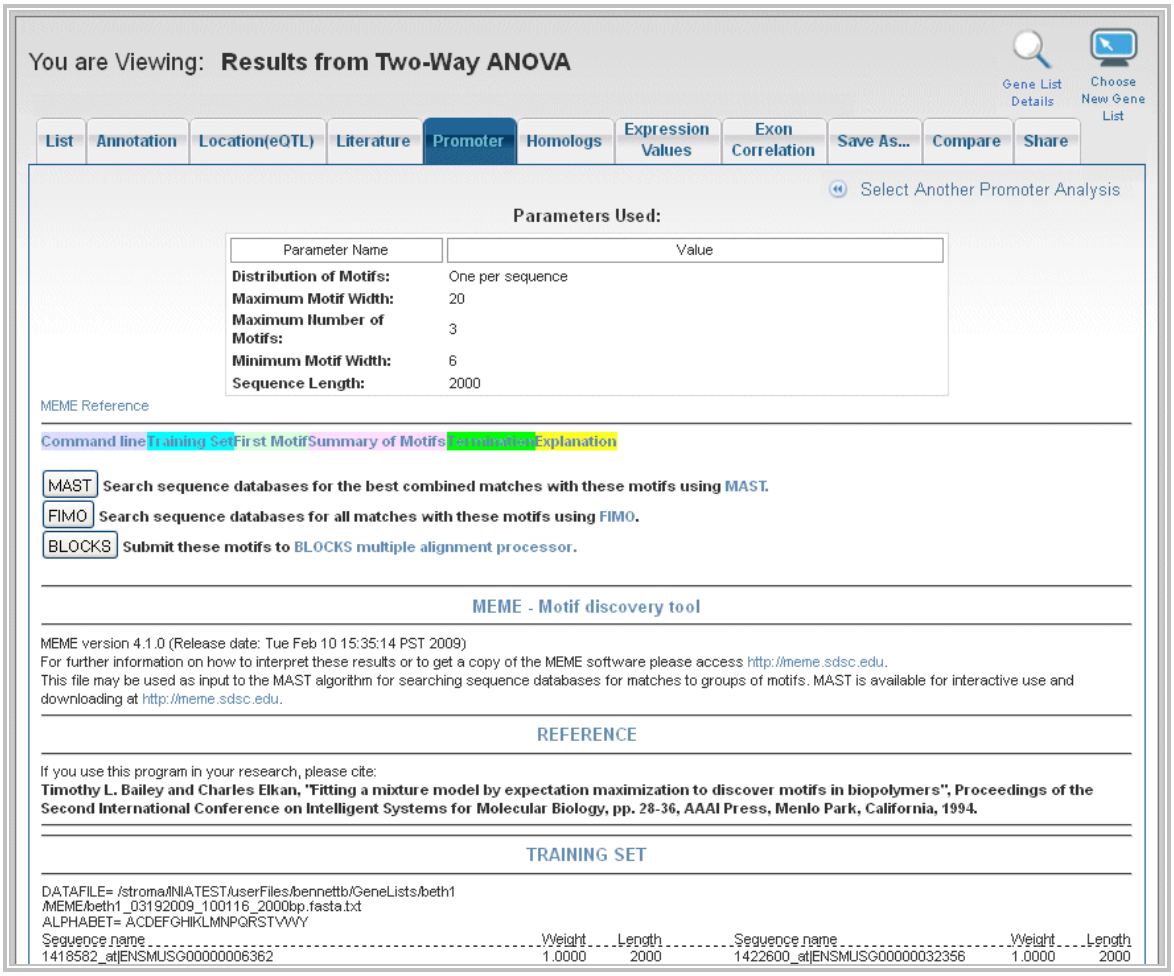
Column Name	Description
TF	The name of the transcription factor.
TF Class	The class of transcription factors to which the transcription factor belongs.
TF SuperGroup	The taxonomic supergroup to which this transcription factor belongs.
IC	The information content or specificity of this TFBS profile's position weight matrix.
Background Gene Hits	The number of genes in the background set for which this TFBS was predicted within the conserved non-coding regions.
Background Gene Non-Hits	The number of genes in the background set for which this TFBS was NOT predicted within the conserved non-coding regions.
Target Gene Hits	The number of genes in the included target set for which this TFBS was predicted within the conserved non-coding regions.
Target Gene Non-Hits	The number of genes in the included target set for which this TFBS was NOT predicted within the conserved non-coding regions.
Background TFBS Hits	The number of times this TFBS was detected within the conserved non-coding regions of the background set of genes.
Background TFBS rate	The rate of occurrence of this TFBS within the conserved non-coding regions of the background set of genes. The rate is equal to the number of times the site was predicted (background hits) multiplied by the TFBS profile, divided by the total number of nucleotides in the conserved non-coding regions of the background gene set.
Target TFBS hits	The number of times this TFBS was detected within the conserved non-coding regions of the target set of genes.
Target TFBS rate	The rate of occurrence of this TFBS within the conserved non-coding regions of the included target set of genes. The rate is equal to the number of times the site was predicted (target hits) multiplied by the TFBS profile, divided by the total number of nucleotides in the conserved non-coding regions of the included target gene set.

Z-score	The likelihood that the number of TFBS nucleotides detected for the included target genes is significant as compared with the number of TFBS nucleotides detected for the background set. Z-score is expressed in units of magnitude of the standard deviation.
Fisher score	The probability that the number of hits vs. non-hits for the included target genes could have occurred by random chance based on the hits vs. non-hits for the background set.
P-value	The probability that the number of hits vs. non-hits for the included target genes could have occurred by random chance based on the hits vs. non-hits for the background set.

1. Click a TF ID to open the JASPER website for that transcription factor.
2. Click a link in the **Target Gene Hits** column to view a list of associated genes for that link. The *Associated Genes* page displays.
3. Click a link in the **Target TFBS Hits** column to view a list of associated genes for that link. The *Associated Genes* page displays.

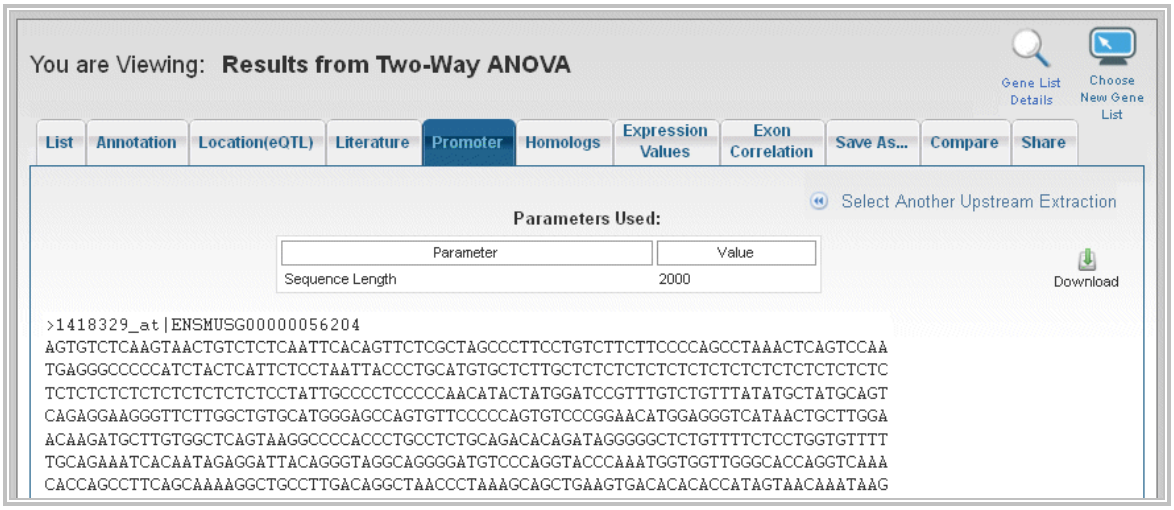
Viewing MEME Results

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view MEME results.
4. Click the **Promoter** tab. The *Promoter* page displays with results organized into tables by the type of analysis performed.
5. Click a row in the *MEME Results* table. The *MEME Results*page displays. An explanation of the MEME results is located at the bottom of the results.



Viewing Upstream Sequence Extraction Results

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to perform a literature search
4. Click the **Promoter** tab. The *Promoter* page displays with results organized into tables by the type of analysis performed.
5. Click a row in the *Upstream Sequence Extraction Results* table. The *Upstream Sequence Results* page displays.
6. Click the **Download** icon to download the results.



Homologs Overview

Homologous genes demonstrate high sequence similarity and can demonstrate similarity in function. Homologous sequences (genes) can be divided into two groups: Orthologs and Paralogs. Homologous sequences (genes) in two different species originating from a common ancestor are know as Orthologs. Duplication of a homologous sequence in a given species results in Paralogous sequences with a different chromosomal location.

The Homolog tab in the PhenoGen website allows you to obtain information regarding chromosomal locations, for genes in a given gene list, in other genomes. For example, you can obtain the chromosomal location for a list of different genes in the mouse genome as well as the chromosomal location for the known homologous genes in the rat and human genomes.

Viewing Homologs

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view homologs.
4. Click the **Homolog** tab. The *Homologous Genes* page displays a table that shows:
    - Gene Identifier
    - HomoloGene ID (NCBI)
    - Gene Symbol in other species
    - Homolog species, identifier, and chromosomal location
5. Click a link in any cell to open the website for your selection.
6. Click the **Download** icon to download the information in the table.

Viewing Pathways

Signaling Pathway Impact Analysis (SPIA) is a tool for analyzing the hypothetical impact of differences in transcription levels of a set of genes on signaling pathways defined by KEGG (Ogata et al 1999). The input is a set of gene identifiers and their fold changes. The **Pathways** tab only displays when a full change or correlation coefficient is available.

The influence of differences in transcription of the input gene set on a particular pathway is assessed using two measures:

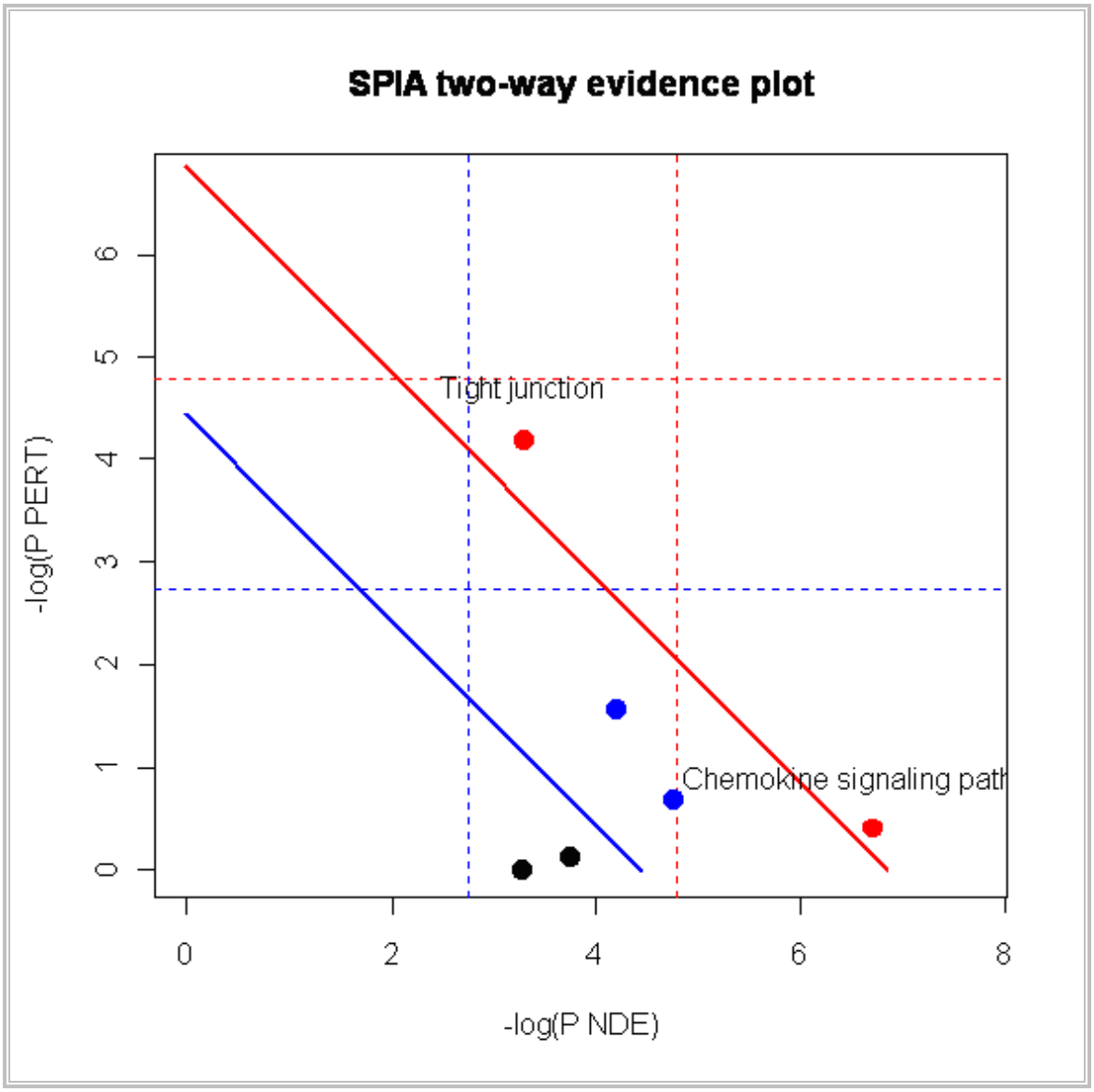
1. **NDE**: The number of genes from the set that are in the pathway.
2. **PERT**: The possible perturbation of the pathway due to changes in expression of genes from the input gene set, that is, the influence of the input genes on the pathway based on their position in the pathway and their magnitude of change.

For each pathway, a p-value is calculated for each of these measures. P_{NDE} is the p-value associated with an enrichment test (i.e., is the number of differentially expressed genes in the given pathway more than one would expected by chance). The values for P_{NDE} are calculated using the assumption that NDE follows a hyper-geometric distribution.

The second probability, P_{PERT}, is calculated based on the estimated amount of perturbation in each pathway due to the differential expression of the input genet set. Each pathway is represented as a network, with genes/proteins for nodes and directed edges indicating interactions between them. The perturbation of the pathway caused by each gene/protein is calculated using the number of genes/proteins it influences (either activates or suppresses) and its magnitude of change.

SPIA takes a table of differentially expressed genes and their fold changes as input and returns a table of signaling pathways containing at least one of the genes on the list. This table summarizes the impact of the differentially expressed genes on each pathway and contains links to images from the KEGG pathways site and to other summary information.

The output also includes a summary plot, where each pathway containing at least one gene from the input list is plotted to its (-log transformed) values for P_{NDE} and P_{PERT}. The plot indicates where the most impacted pathways lie, with respect to two statistical thresholds. The first is the family-wise error rate, indicated by the solid red line in the plot. The second is the false discovery rate, indicated by the solid blue line in the plot.



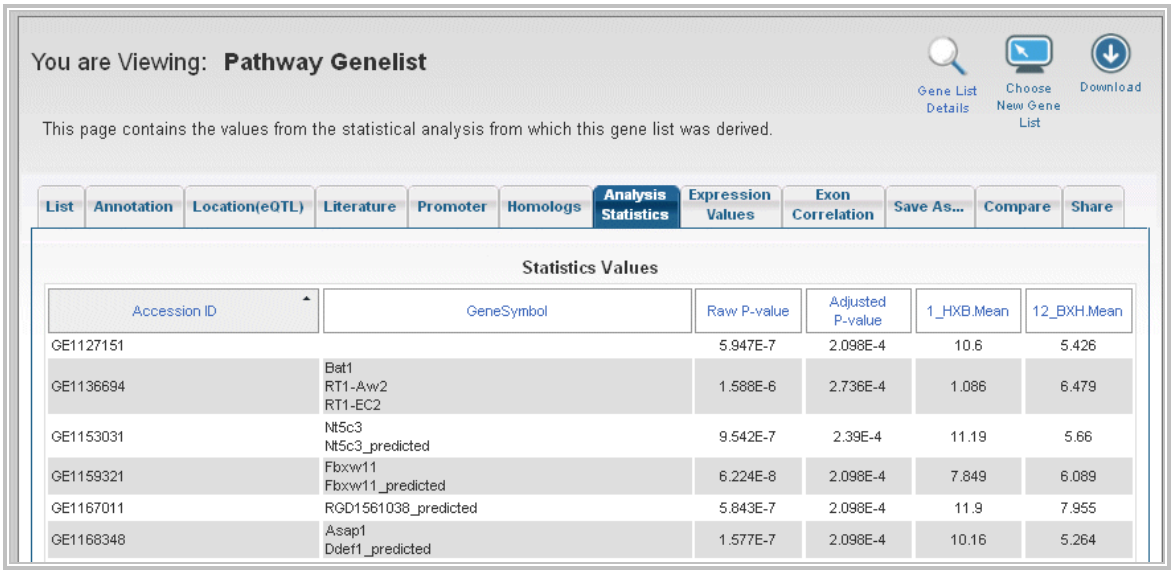
References

1. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27(1):29-34.
2. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R (2009). A novel signaling pathway impact analysis. Bioinformatics 25(1):75-82.

Viewing Analysis Statistics

The **Analysis Statistics** tab only displays if your gene list was derived from the analysis of a dataset.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view analysis statistics.
4. Click the **Analysis Statistics** tab. The *Statistics Values* table displays.



The tables displays gene identifiers, gene symbols, the raw p-values and the adjusted p-values. Depending on the number of groups and the type of analysis used, it may also display group means, F-statistic, mean intensity, correlation coefficient, difference in log base 2 intensity, t-statistic, or parameter estimates.

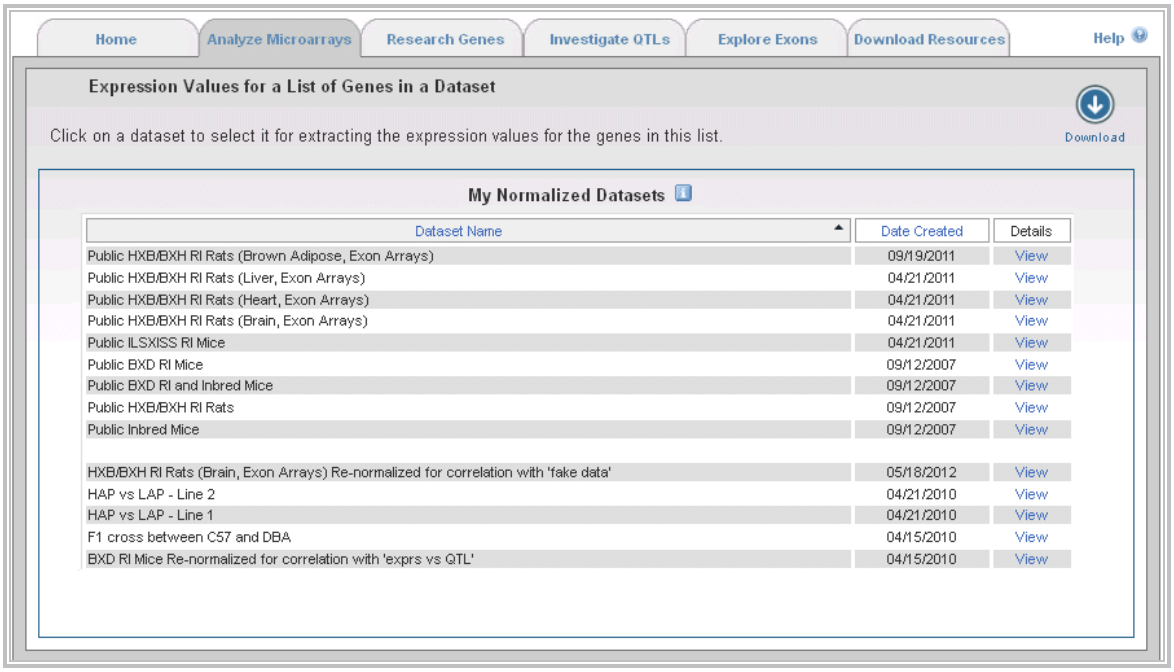
Viewing Gene Expression Data

The *View Gene Expression* page allows you to obtain gene expression intensity values for a gene or gene list from a normalized dataset, or specify a pre-created gene list. The results table displays group-level information for each gene, such as group means and group standard error.

You can view gene expression data from the Analyze Microarrays or the Research Genes tabs:

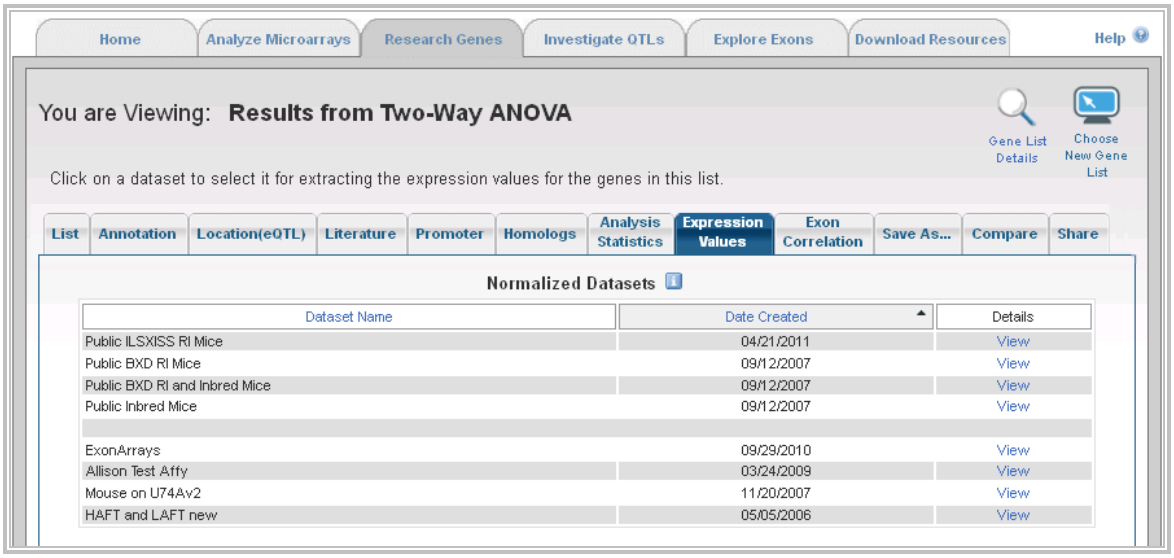
Analyze Microarrays tab:

1. Click the **Analyze Microarrays** tab. The *Analyze Microarrays* tab displays.
2. Click **View expression values for a list of genes in a dataset** in the *What would you like to do* section. The *Expression Values* page displays.



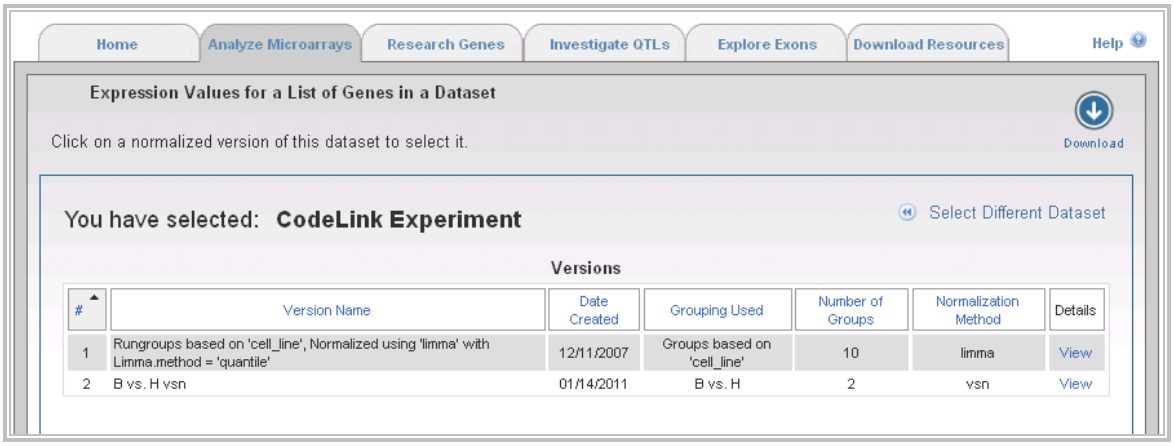
Research Genes tab:

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view gene expression data.
4. Click the **Expression Values** tab. The *Expression Values* page displays.

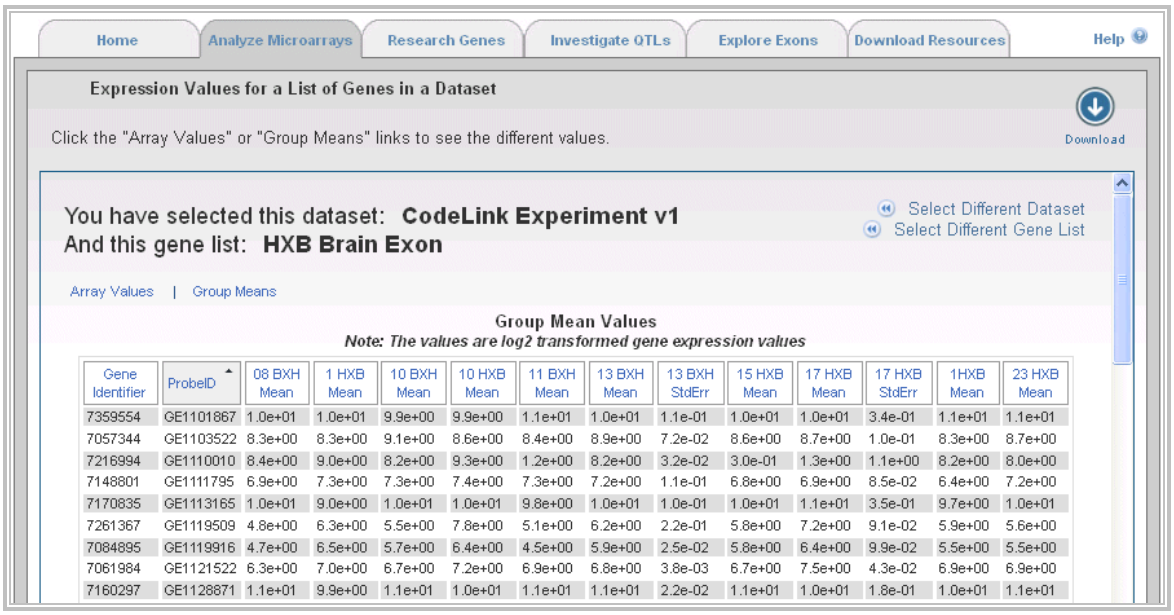


On the *Expression Values* page:

1. **OPTIONAL:** Click the **View** link in the **Details** column to view dataset details such as name, description, organism, arrays in dataset, and more. See "Viewing Dataset Details" for more information.
2. Click the **dataset** for which you want to see gene expression data. A page displays the normalized versions of that dataset.



3. Click the **normalized version** for which you want to see gene expression data. A page displays gene lists for that version.
4. Click the **gene list** for which you want to see gene expression data. The *View Gene Expression* page displays.



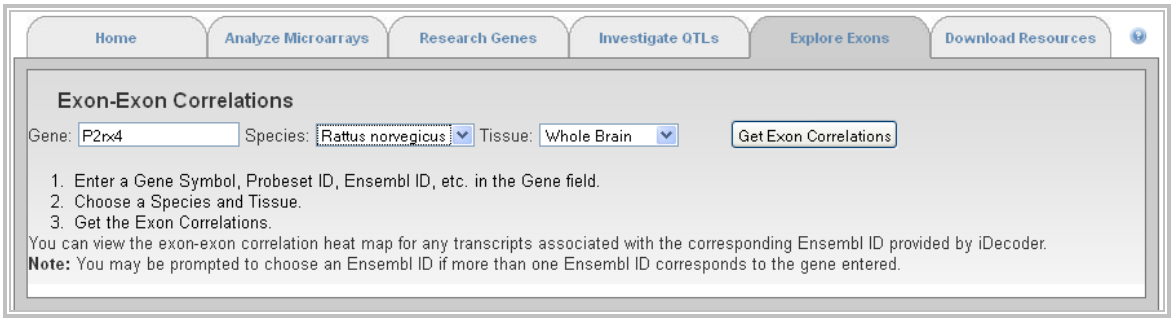
5. You can:
   - Click the **Array Values** link to view the individual array values.
   - Click the **Group Means** link to view the group mean values (this is the default view).
   - Click **Download** to save the group means as well as the individual array values. Follow the instructions that display as you open or save the files. These instructions vary depending on your Internet browser (e.g., Internet Explorer, Firefox, Safari).

Viewing Exon-level Correlations

You can create an exon correlation heatmap for a specific gene, species, and tissue. The data is retrieved from multiple databases, so generating the initial graphics may take a few moments.You can display the heat-maps for two transcripts of a gene side-by-side to determine the transcript that best fits the expression cor-relation patterns.

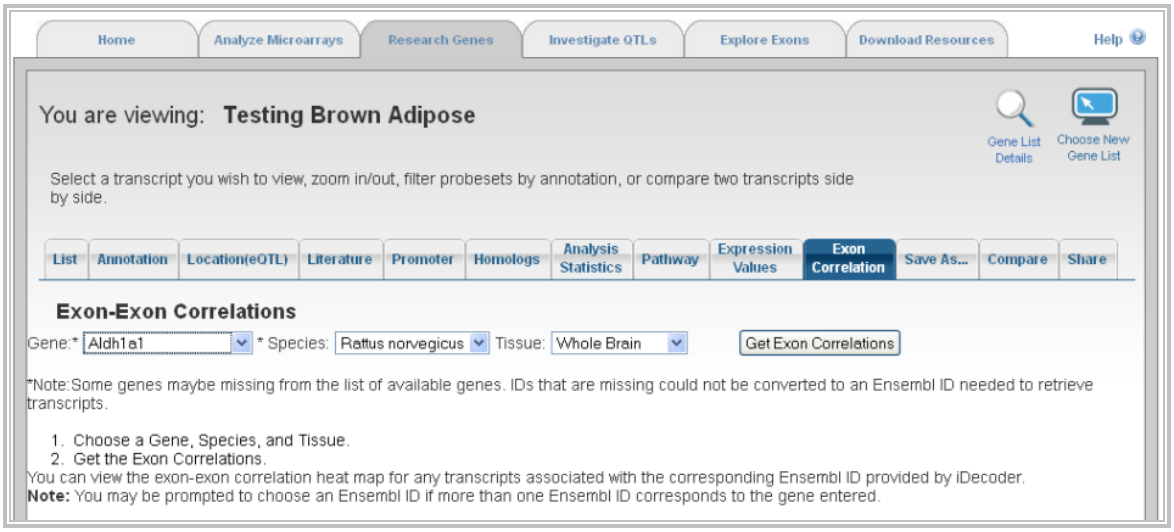
Explore Exons tab:

1. Click the **Explore Exons** tab. The *Explore Exons* tab displays.
2. Click **View exon-level information for a gene** in the *What would you like to do* section. The *Exon-Exon Correlations* page displays.



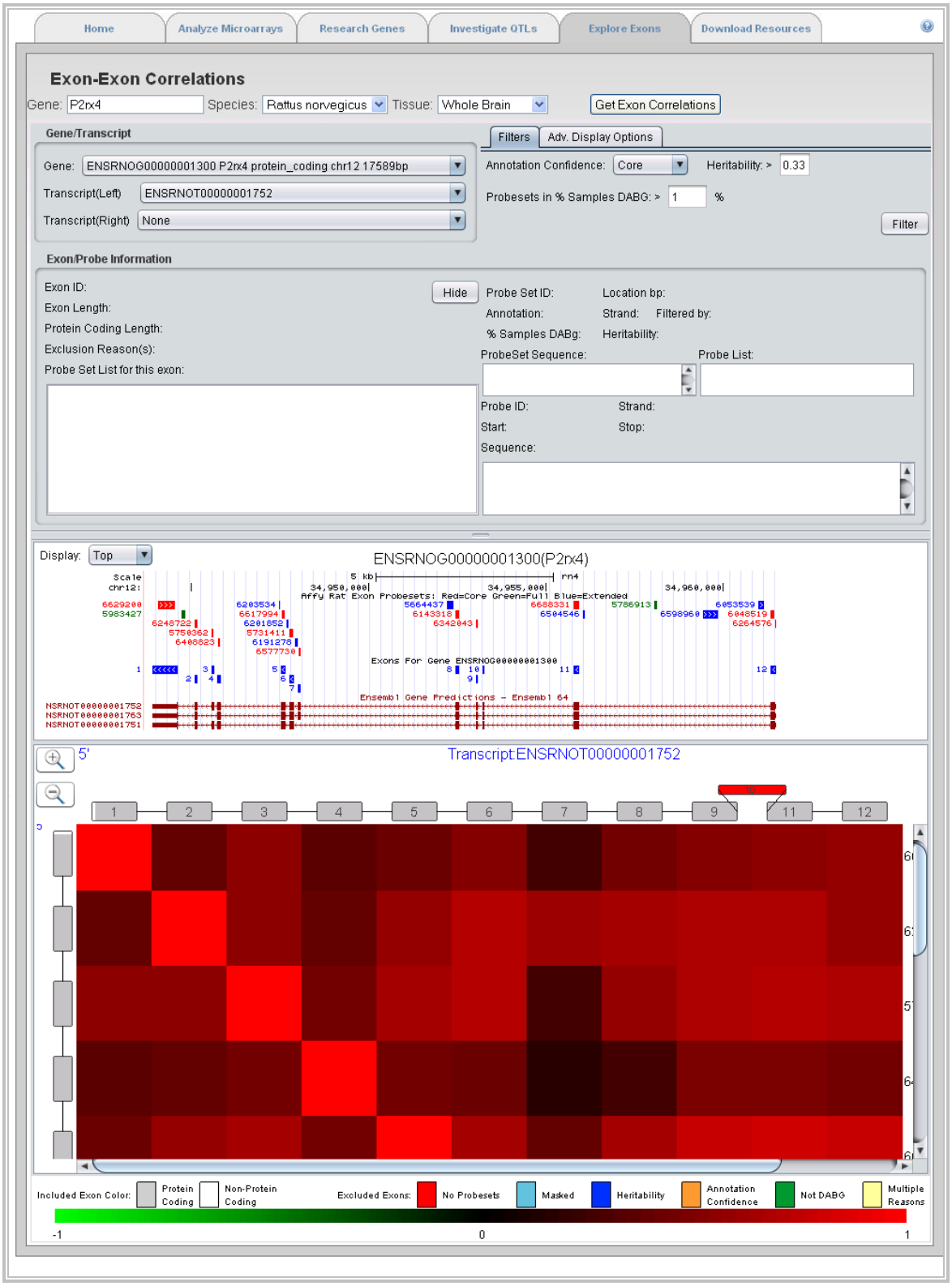
Research Genes tab:

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view gene expression data.
4. Click the **Exon Correlation**tab. The *Exon-Exon Correlations* page displays.



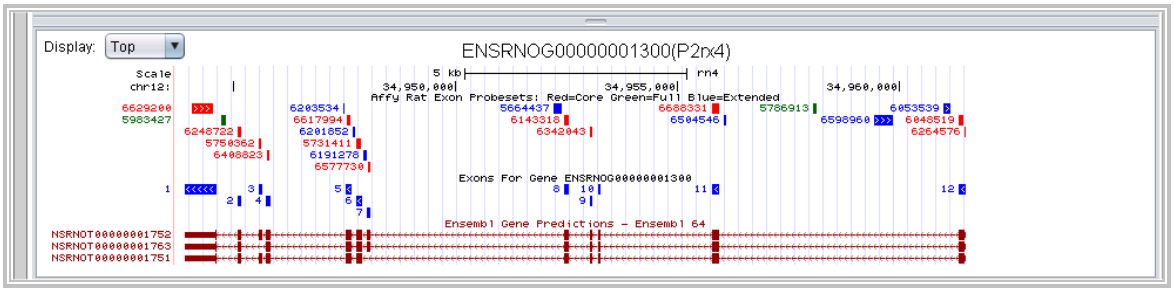
On the Exon-Exon Correlations page:

1. Enter the **Gene** for which you want to create an exon correlation. On the *Exon Correlation* tab, the available genes for the selected Gene List are available in a drop-down list.
2. Select the **Species** and the **Tissue** to which this correlation pertains.
3. Click **Get Exon Correlations**. The exon correlations display.



4. **OPTIONAL:** Display two heat maps by selecting both a Left and Right Transcript in the In the *Gene/Transcript* section at the top left.
5. **OPTIONAL:** Hover over the schematic on top of the heatmap or click on individual exons in the schematic to find out more information about individual exons and the probesets that align to them.

By default, the first graphic, which is generated through the University of California Santa Cruz Genome Browser (http://genome.ucsc.edu/), displays the probesets from the exon array and the annotated Ensembl transcripts (i.e., isoforms) for the input gene.



The exons are color-coded based on their level of annotation according to Affymetrix. The red probesets are included in the core set of probesets and indicate high confidence in the annotation. The blue probesets are included in the extended set of probesets according to Affymetrix, and the green probesets are from the full set of probesets and indicate the least amount of confidence in the annotation. The next track assigns a number to each annotated exon of the gene to aid in interpretation of the heatmap that follows.

By default, the second graphic is a heatmap that displays the correlations among exons of a given transcript of the gene.

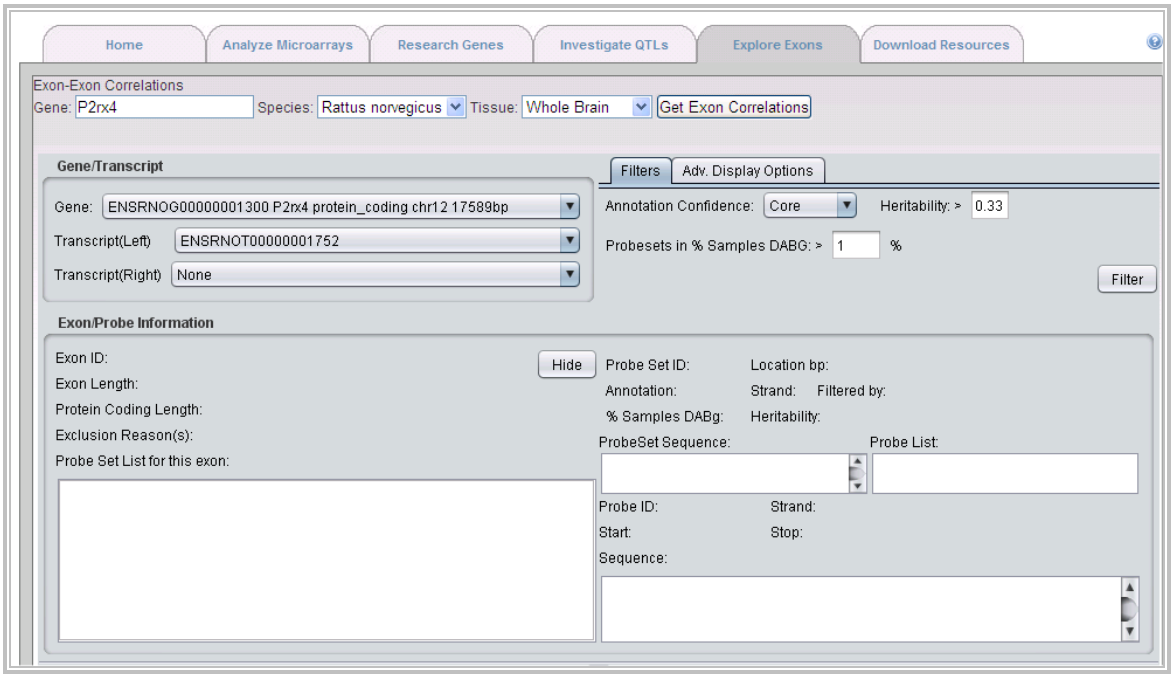
Each row and column represents a probeset that passed the filtering criteria specified at the top right side of the page. Along the top of the heatmap is a schematic of the exon structure of the transcript. The boxes represent exons of the transcript and are numbered, as in the top figure. Grey boxes are exons that are probed by the array, met the filtering criteria, and are displayed in the heatmap. Exons that do not satisfy these criteria are colored based on the reason why they are not included in the heatmap.

- **No Probesets (Red)** indicates that the Affymetrix array does not contain any probesets that target this exon.
- **Masked (Light Blue)** indicates that the probeset was removed due to poor intergrity of its probes (i.e., the probes aligned to many regions of the genome or targeted a region of the genome that harbored a known SNP).
- **Heritability (Blue)** indicates that the probeset had a low (according to filtering criteria) heritability in the RI panel.

- **Annotation Confidence (Orange)** indicates that probesets that align to this exon are not included due the annotation confidence criteria selected for filtering.
- **Not DABG (Green)** indicates that the probesets that align to this exon are not expressed above background according to the DABG criteria selected for filtering.
- **Multiple Reasons (Yellow)** indicates that the exon was filtered out for more that one reason, e.g., low heritability and not DABG

In the top right section, choose the probesets to display for the heatmap based on the:

- Annotation Confidence from Affymetrix (Core, Extended, or Full).
- Heritability of the probeset's expression in the RI panel.
- Proportion of samples with expression values above background (Detected Above Background–DABG).

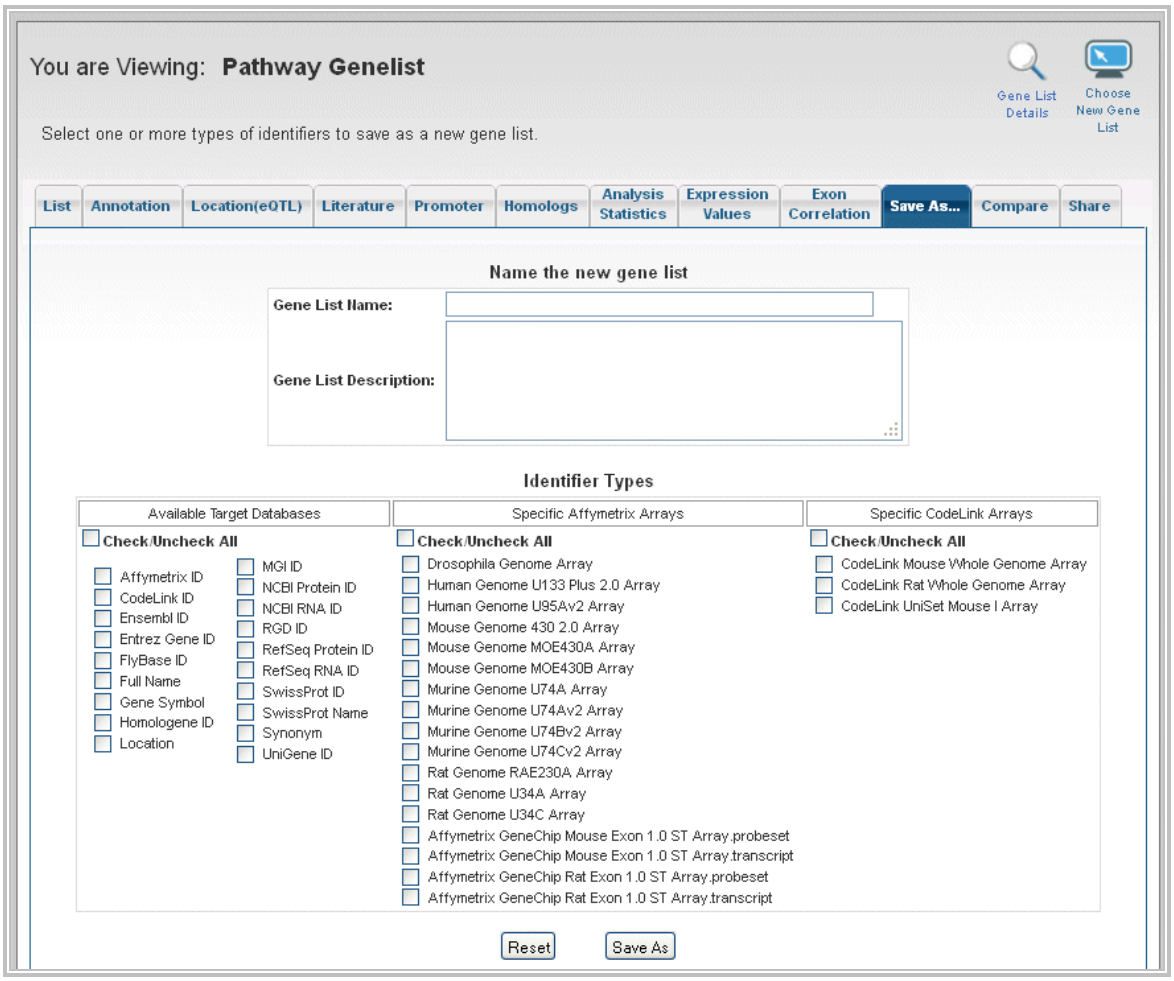


There are also advanced display options that allow you to include probesets that align to an intronic region of the gene or align to the opposite strand from which the gene is coded.

Saving a Gene List as Other Identifiers

You can save a gene list using different identifier types.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list that you want to edit and save as a new gene list.
4. Click the **Save As** tab. The *Save As* page displays.



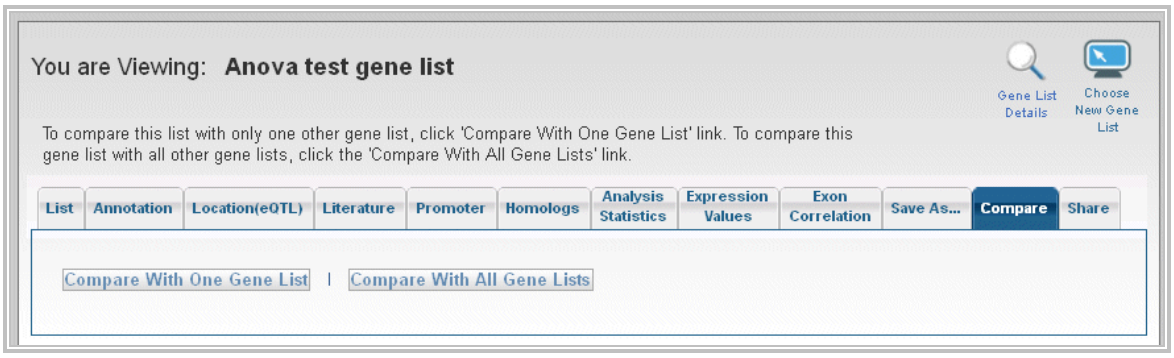
5. Enter a new **Gene List Name**.
6. Enter the **Gene List Description**.
7. Choose **Identifier Types** from the Available Target Databases, Specific Affymetrix Arrays, and Specific CodeLink Arrays.
8. Click **Save As**. The gene list is copied and saved with the new name, description,and identifiers from the selected database. Click **Reset** to return all the fields to their original values.

See "Viewing Gene Lists" for instructions on viewing the new gene list.

Comparing Gene Lists

The *Compare Gene Lists* page allows you to compare gene lists.

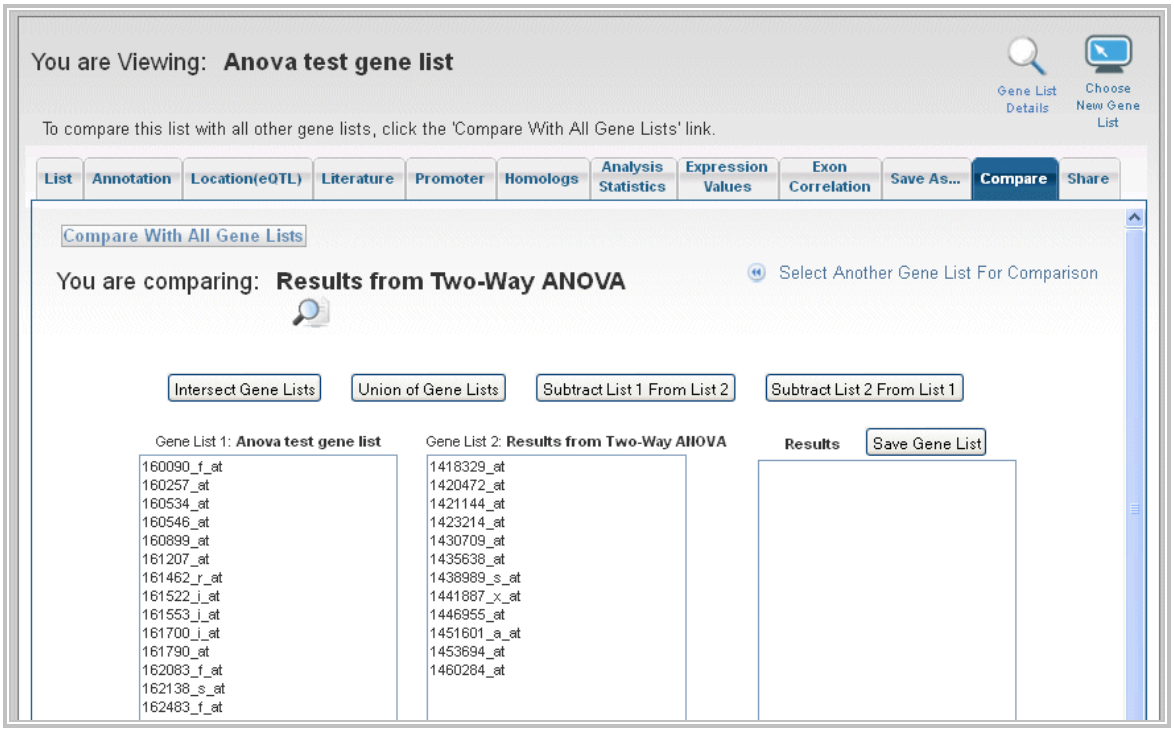
1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list that you want to compare with other gene lists.
4. Click the **Compare** tab.



5. Choose one of the following:
   - **Compare With One Gene List**. A table of gene lists for comparison displays.
   - **Compare With All Gene Lists**.

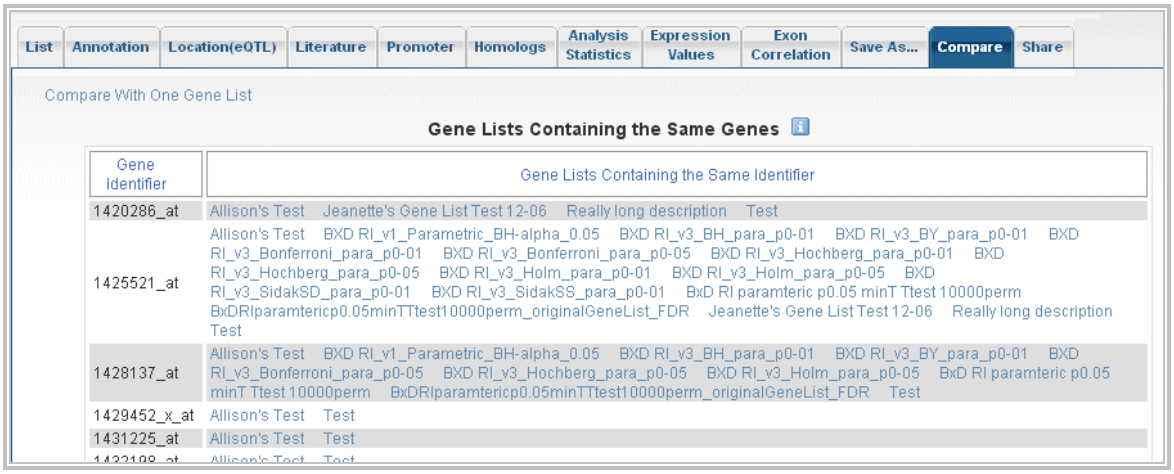
If you choose *Compare With One Gene List:*

6. Choose a gene list from the table that displays.



7. Do one of the following:
    - Click **Intersect Gene Lists** to view a list that displays the genes that are in both selected lists.
    - Click **Union of Gene Lists** to combine the two gene lists into one gene list
    - Click **Subtract List 1 from List 2** to remove the genes in list 1 from list 2. The resultant gene list displays the genes that are only in list 2.
    - Click **Subtract List 2 from List 1** to remove the genes in list 2 from list 1. The resultant gene list displays the genes that are only in list 1.
8. Click **Save Gene List** if you want to save the resulting gene list.

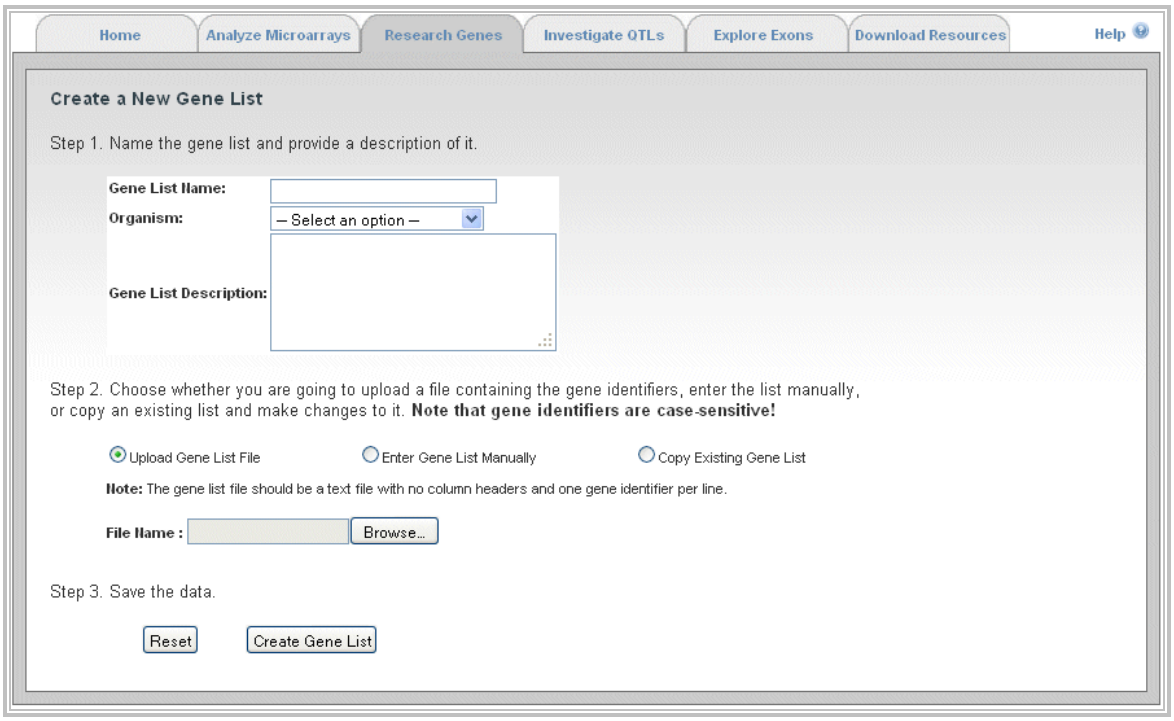
If you choose *Compare With All Gene Lists*, a table displays all the gene lists that contain the same genes as the list you chose to compare with. Click the name of a gene list to view it.



Uploading a Gene List

You can upload an existing gene list to the PhenoGen website for analysis. Your gene list must be a text file and must contain only one gene identifier per line or it will not upload correctly. When you upload gene list data, you can give other users permission to see the gene list.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Upload or create a new gene list** in the *What would you like to do* section. The *Create a New Gene List* page displays.



3. Enter the **Gene List Name**.
4. Select the **Organism** from the drop-down list.
5. Enter the **Gene List Description**.
6. Make sure **Upload Gene List File** is selected.
7. Click **Browse**, and follow the instructions to select the gene list you want to upload.
8. Click **Create Gene List**. The new gene list is created.

Downloading a Gene List

You can download a gene list from the PhenoGen website to analyze it using your own tools or to distribute it to others.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the **Download** icon beside the gene list you want to download.
4. Click the checkbox next to the gene list(s) you want to download. The second and third options only display if the gene list was derived from a microarray analysis.

5. Click **Download**.
6. Choose to open or save the files, and follow the instructions that display. These instructions vary depending on your Internet browser (e.g., Internet Explorer, Firefox, Safari, etc.).

Deleting a Gene List

If you no longer need a gene list, you can delete it.

Important! Any references to the gene list, such as literature searches and promoter analysis results, are deleted when the gene list is deleted.

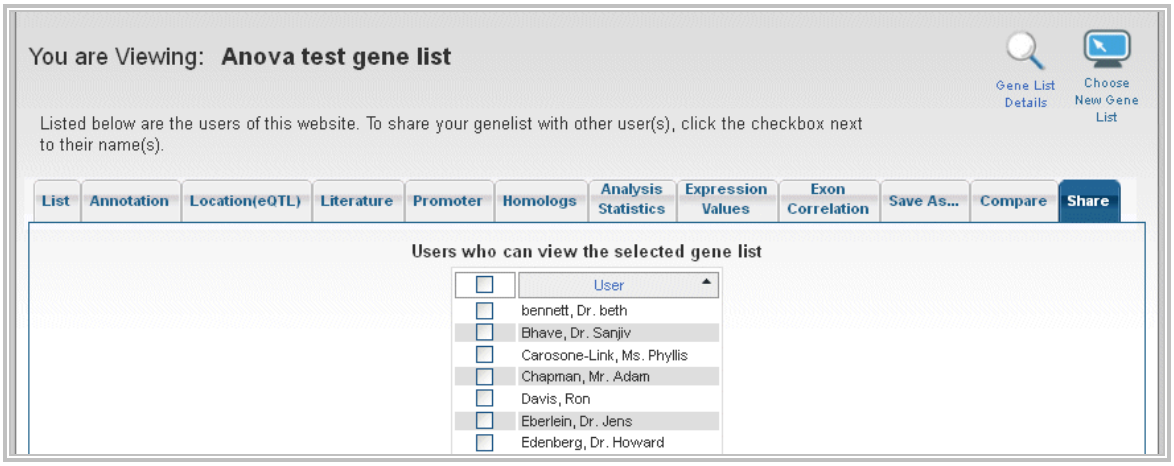
1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the **Delete** icon beside the gene list you want to delete.
4. Review the data that is linked to the gene list.
5. Click **Delete Gene List**. The gene list is deleted.

Sharing a Gene List

You can view the users who have access to your gene lists and grant users access to a gene list that you own. You cannot share a gene list that you do not own.

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view user access.
4. Click the **Share** tab.

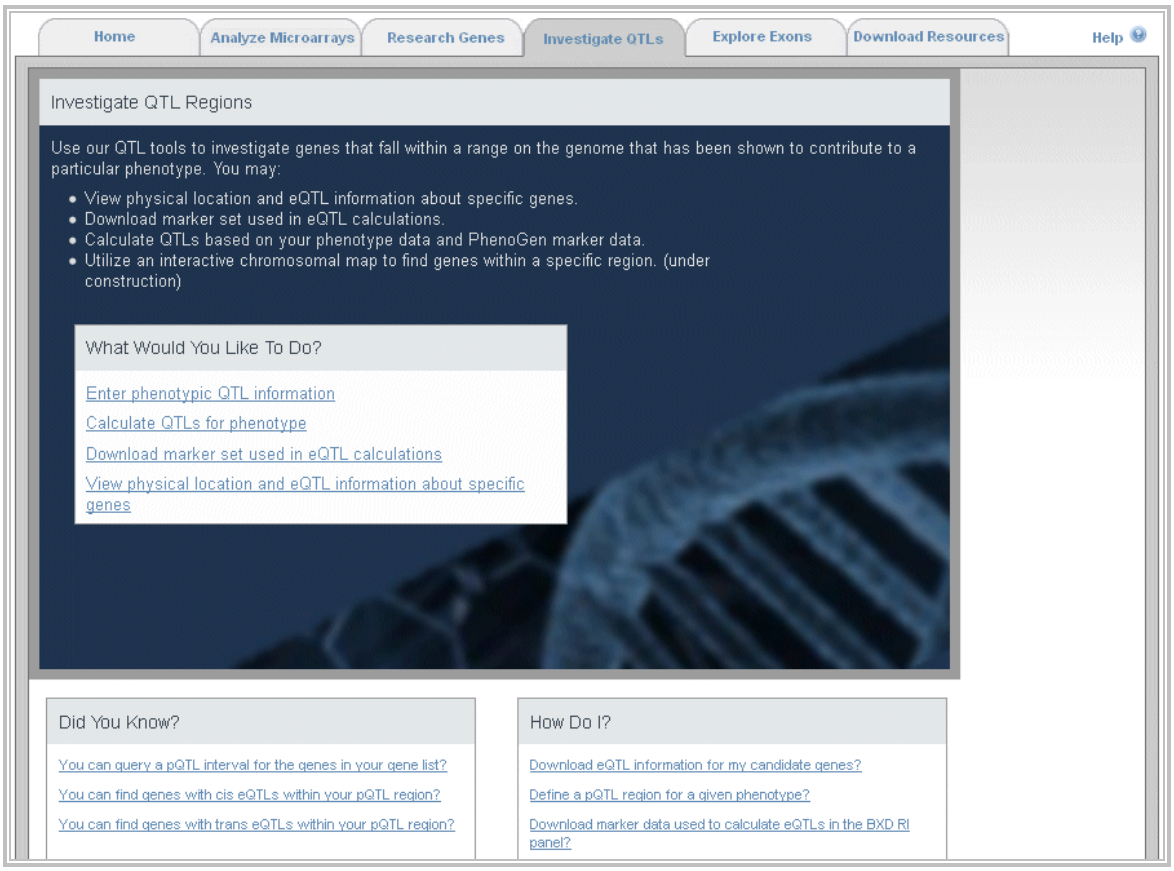
Note: If you are the owner of the gene list, you can allow other users to view the gene list.



5. Select the users to whom you want to give view permissions for your gene list.
6. Click **Update Gene List Users** at the bottom of the window.

Investigating QTL Regions

The **Investigate QTL Regions** tab allows you to use QTL tools to assess whether the genomic location of any of the genes in a gene list fall within the QTL regions for the phenotypes of your choice. Information about the phenotypic QTLs can be obtained from MGI (mouse QTLs) or RGD (rat and mouse QTLs).



Choose an option to get started:

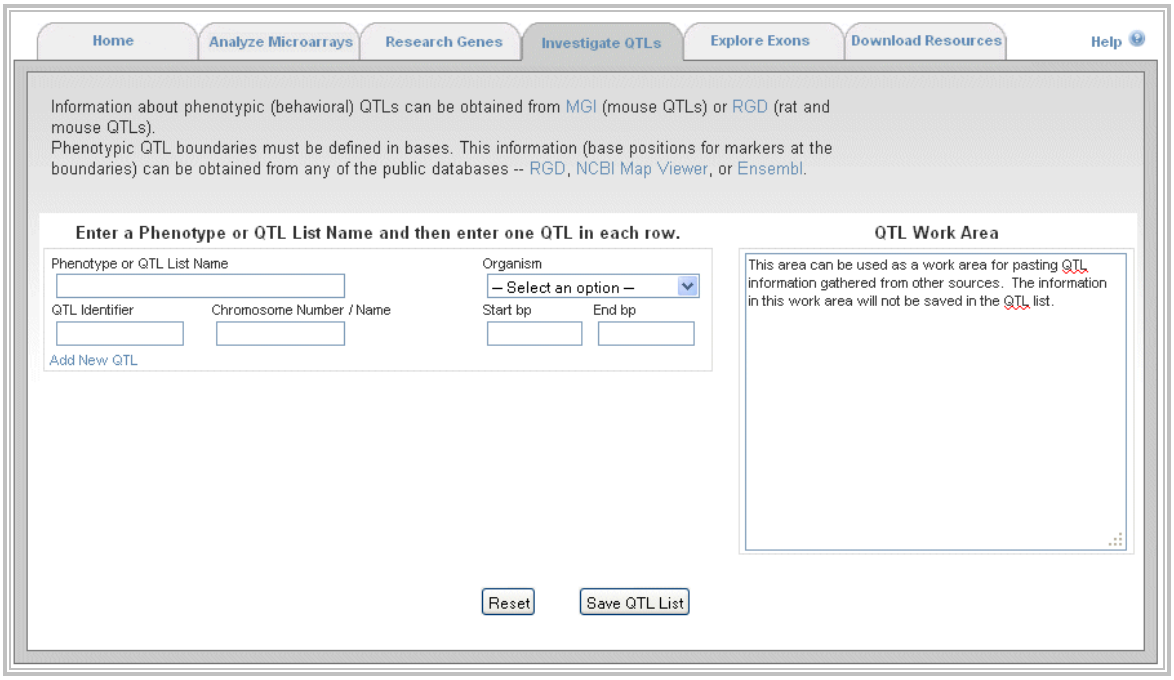
- Enter phenotypic QTL information. See "Entering Phenotypic QTLs".
- Calculate QTLs for Phenotype. See "Calculating QTLs for Phenotype".
- Download marker set used in eQTL calculations. See "Downloading eQTL Marker Sets".
- View physical location and eQTL information about specific genes. See "Viewing Location and eQTL".

Entering Phenotypic QTLs

You can enter phenotypic data to create a list of QTLs for a phenotype. For example, you could define a set of QTLs for alcohol preference.

1. Click the **Investigate QTL Regions** tab. The *Investigate QTL Regions* page displays.
2. Click **Enter Phenotypic QTL information** in the *What would you like to do* section. The *Entering Phenotypic QTLs* page displays.
3. Enter the **Phenotype** or **QTL List Name**. This field is referenced on the *Viewing Location and eQTL* page.
4. Select the organism to which this list pertains.
5. Enter the **QTL Identifier**, **Chromosome Number**, and **Start** and **End** base pair (bp) positions.

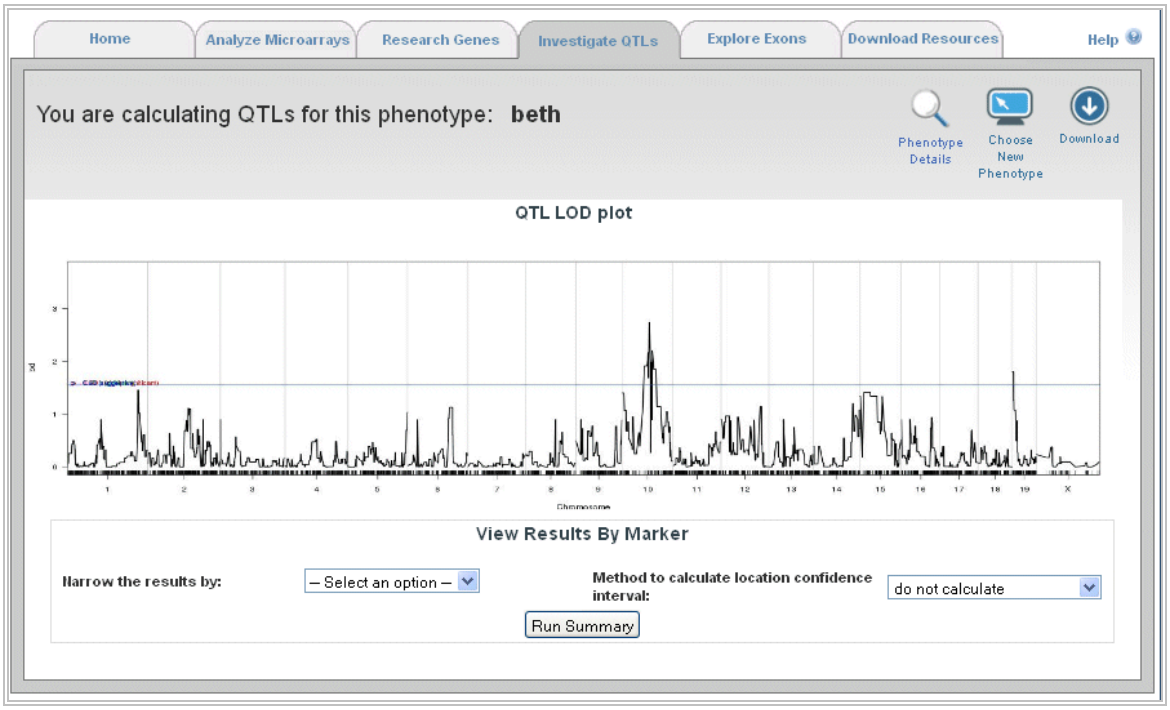
6. **OPTIONAL:** Click the **Add New QTL** link to create more than one QTL range for a particular phenotype.
7. Click **Save QTL List**.



Calculating QTLs for Phenotype

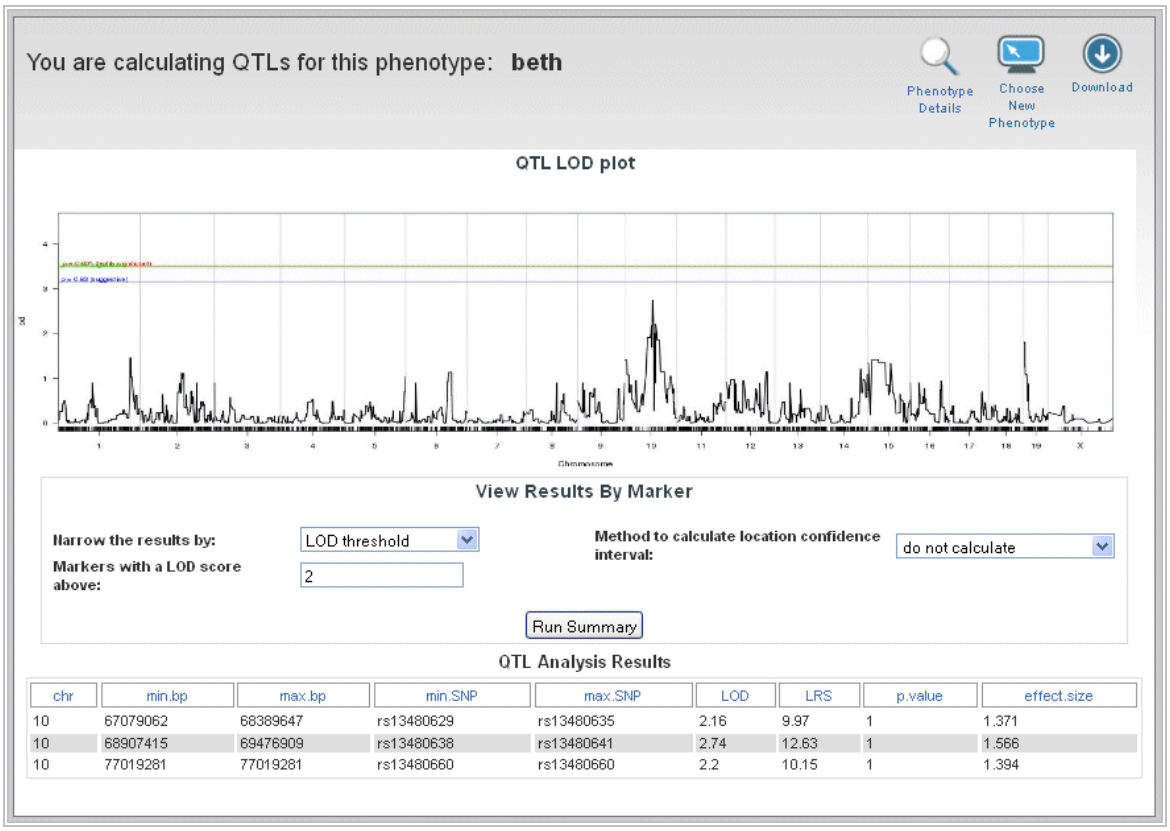
You can calculate QTL for uploaded phenotype data using pre-compiled marker sets. The website supports phenotypic QTL analysis for BXD recombinant inbred mice and HXB/BXH recombinant inbred rats.

1. Click the **Investigate QTL Regions** tab. The Investigate QTL Regions page displays.
2. Click **Calculate QTLs for Phenotype** in the *What would you like to do* section. The Calculate QTLs for Phenotype page displays.
3. Choose the Marker Set data: **Public BXD RI Mice**, **Public HXB/BXH RI Rats**, or **Public ILSXISS RI Mice**.
4. **OPTIONAL:** Click **Enter New Phenotype Data**. You can upload, manually enter, or copy phenotype data. Click **Enter Variance Values** to enter variances for your groups. See "Using Phenotype Data in Correlation Analysis" for upload, enter, copy, and delete instructions.
5. Click a row in the Phenotype Values table.
6. Choose **Yes** or **No** to weight the analysis based on variance.
7. Enter the **number of permutations** between 0 and 1 000 000. Zero indicates that empirical p-values are not calculated.
8. Click **Run**. A LOD plot displays.
9. Choose an option to narrow the results: **LOD threshold**, **p-value threshold**, **genomic location**, and enter associated criteria.
10. Choose a **method to calculate confidence interval** (see descriptions in the *Summary of Results* section).
11. Click **Run Summary**. The results that satisfy your criteria display.



QTL Calculation

Due to the large marker density of the two available marker sets, the QTL calculation is done using a simple marker regression. LOD (log base 10 odds ratio) scores are calculated for each marker and displayed on the LOD plot. When strain variances or standard deviations are available you can weight the marker regression by the inverse of the strain variance, to give more weight to strain means with lower variance.



Marker Sets

Markers for BXD Recombinant Inbred Mice

Markers for the BXD recombinant inbred panel were retrieved from the Wellcome-CTC Mouse Strain SNP Genotype Set (http://www.well.ox.ac.uk/mouse/INBREDS/). SNPs were retained if their marker ID could be matched with a SNP in the Ensembl 56 dbSNP128 mouse data retrieved through BioMart (http://www.ensembl.org/biomart/index.html) and if the SNP was also give a valid position in the mouse genome. In addition, individual SNPs were dropped if they did not differ between the two parental strains (C57BL/6J and DBA/2J) and if they had unknown genotypes for all BXD RI strains.

11 SNPs were removed because they had more than three strains with double recombinant genotypes. A double recombinant genotype occurs when a SNP between two SNPs that are less than 1 Mb has a different genotype for a strain than the SNPs on either side. Often these double recombinants represent a genotyping error rather than two instances of recombination so close together.

As an additional quality control measure, individual SNPs were evaluated for segregation distortion. This measure looks at the proportion of subjects/samples in each genotype category and assesses the probability of this distribution. In the case of recombinant inbred strains, the probability of each genotype is 50%. Two SNPs (rs13459145 and rs13478690) were dropped because of significant segregation distortion ($p < 1.0 X 10e-8$).

Finally, a genomic map was estimated from the remaining SNPs, and if a SNP was estimated to have more than 40 cM between it and another SNP on either side, it was eliminated from the data set (one SNP). These quality control measures resulted in a final dataset with 91 strains including the parental strains and 6,093 SNPs.

Markers for HXB/BXH Recombinant Inbred Rats

Markers for the HXB/BXH recombinant inbred panel were retrieve from the STAR consortium (http://www.snp-star.eu/). SNPs were eliminated if they did not differ between parental strains (SHR/Ola and BN-Lx), if they were not genotype for either parental strain, or if they were heterozygous for either parental strain. If the genotype for a recombinant inbred strain was unknown, but the two surrounding SNPs had the same genotype, the missing genotype was assigned the value of the surrounding SNPs (53% of unknown genotypes were changed).These quality control measures resulted in a final dataset with 34 strains including the parental strains and 13,143 SNPs.

Genome-wide P-Values

A genome-wide p-value is calculated using permutation (Churchill and Doerge 1994) to account for multiple testing across markers. To calculate these p-values, phenotype and genotype associations are permuted and for each permutation, the maximum LOD score is retained. The distribution of these maximum LOD scores then becomes the null distribution from which p-values are calculated. For example the p-value of a LOD score of 3 is equal to the proportion of permutations where the maximum LOD score was greater than 3. Although results for all markers are shown in the LOD plot, only unique strain distribution patterns were used in the calculation of genome-wide p-values.

SNPs have identical strain distribution patterns in inbred panels, and when strains are grouped by genotype for either marker, the same two groups are formed. When two SNPs have identical strain distribution patterns, their LOD score is the same. Identical strain distribution patterns in adjacent markers are generally caused by linkage disequilibrium while identical strain distribution patterns in non-adjacent markers may be a consequence of using only a small number of strains.

Summary of Results

After the QTL calculation, you can limit the results to markers that reach some threshold (p-value or LOD score) or to markers within a certain genomic area. Results are shown on the strain distribution pattern (SDP) level. The minimum and maximum basepair positions on each chromosome are shown for each SDP that meet the criteria.

You can also calculate location confidence intervals for QTLs that met the criteria you specified when you summarized the results. In addition to not calculating confidence intervals at all, three different methods are available: non-parametric bootstrap, LOD support interval, and Bayesian credible interval. Confidence intervals are only calculated for the maximum LOD score on each chromosome.

- The non-parametric bootstrap confidence intervals are calculated using the methods outlined by Visscher et al (1996). One thousand bootstrap samples are taken to calculate intervals, and you can choose your own probability coverage. The bootstrap non-parametric measure is useful when it is possible that more than one QTL is in the area.
- For the LOD support interval, the confidence interval is identified by locating the marker on either side of the maximum LOD scores were the LOD score first drops by the LOD score threshold you specified.
- The Bayesian credible interval is calculated as outlined in Manichaikul et al (2006). This article also compares the three methods and recommends against the bootstrap non-parametric method, but their analysis was done with relatively sparse marker maps on backcross and intercross populations using only one QTL.

References

1. Churchill GA and Doerge RW (1994). Empirical threshold values for quantitative trait mapping. Genetics Nov;138(3):963-71.

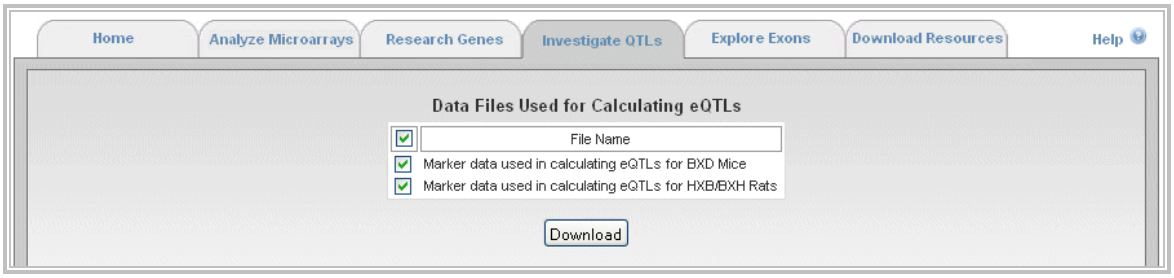
2. Manichaikul A, Dupuis J, Sen S, Broman KW (2006). Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. Genetics 174(1):481-9.

3. Visscher PM, Thompson R, and Haley CS (1996). Confidence intervals in QTL mapping by bootstrapping. Genetics 143(2):1013-20.

Downloading eQTL Marker Sets

You can download the data files used to calculate eQTLs.

1. Click the **Investigate QTL Regions** tab. The *Investigate QTL Regions* page displays.
2. Click **Download marker set used in eQTL calculations** in the *What would you like to do* section. The *Download Marker Sets* page displays.
3. Click the checkbox next to the data file(s) you want to download.
4. Click the **Download** button.



5. Choose to open or save the files, and follow the instructions that display. These instructions vary depending on your Internet browser (e.g., Internet Explorer, Firefox, Safari, etc.).

Viewing Location and eQTL

The location graphic allows you to investigate and use expression QTL (eQTL) datasets. By default, only the physical locations are shown on the graphic. You can click the "Show locations of transcription control in brain" check box to display the locations of transcription control. These green arrows represent the genomic position(s) that control transcription of the candidate genes in the gene list (expression QTLs, or eQTLs).You can also change the significance level for displaying eQTLs

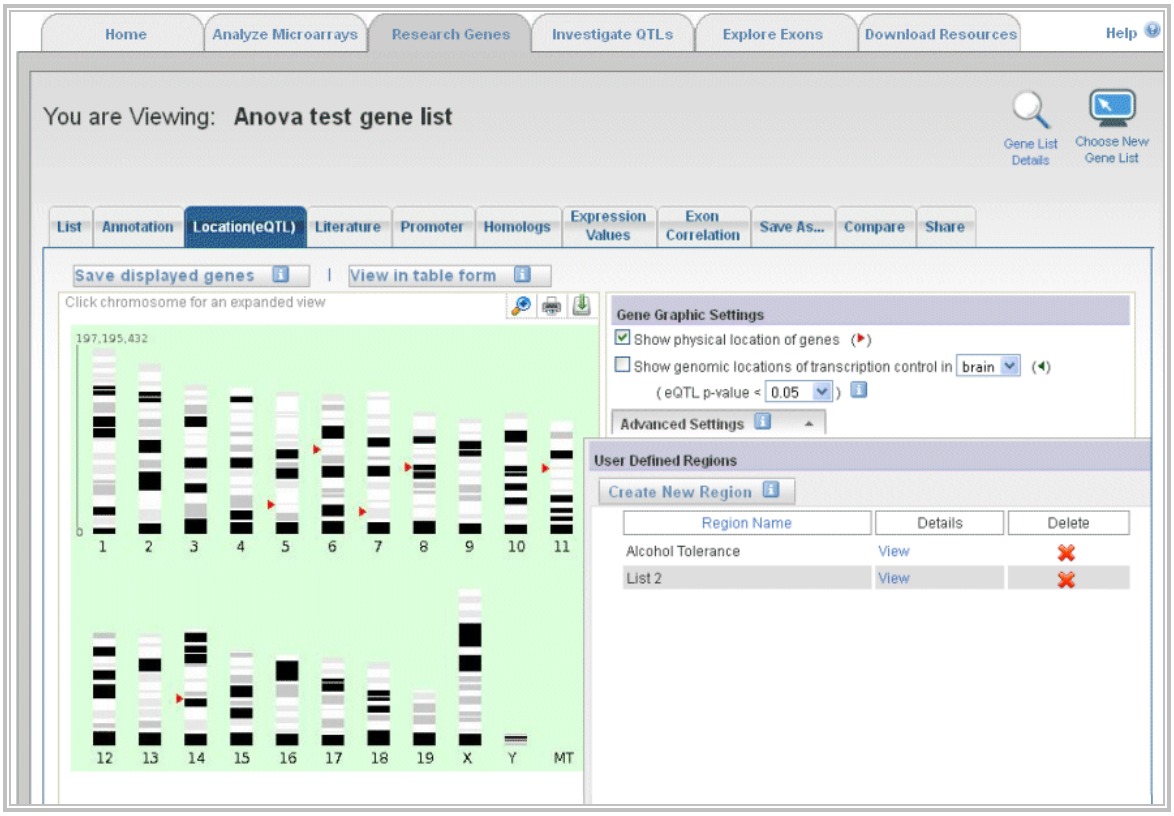
You can view the location and eQTL data for a gene list in two ways:

From the *Research Genes* tab:

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view location and eQTL.
4. Click the **Location** tab. The *Location* page displays.

From the *Investigate QTL Regions* tab:

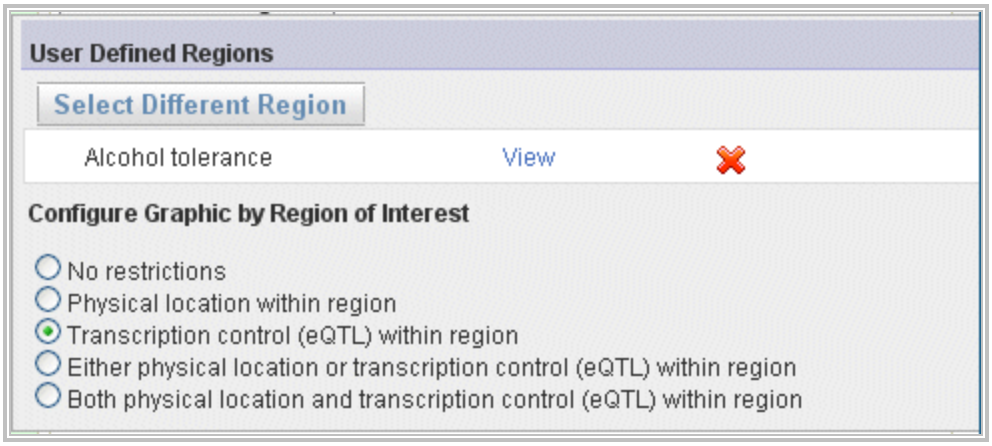
1. Click the **Investigate QTL Regions** tab. The *Investigate QTL Regions* page displays.
2. Click **View physical location and eQTL information about specific genes** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view location and eQTL. The *Location and eQTL* page displays.



The data on the page looks the same, regardless of the way you access the page. You can:

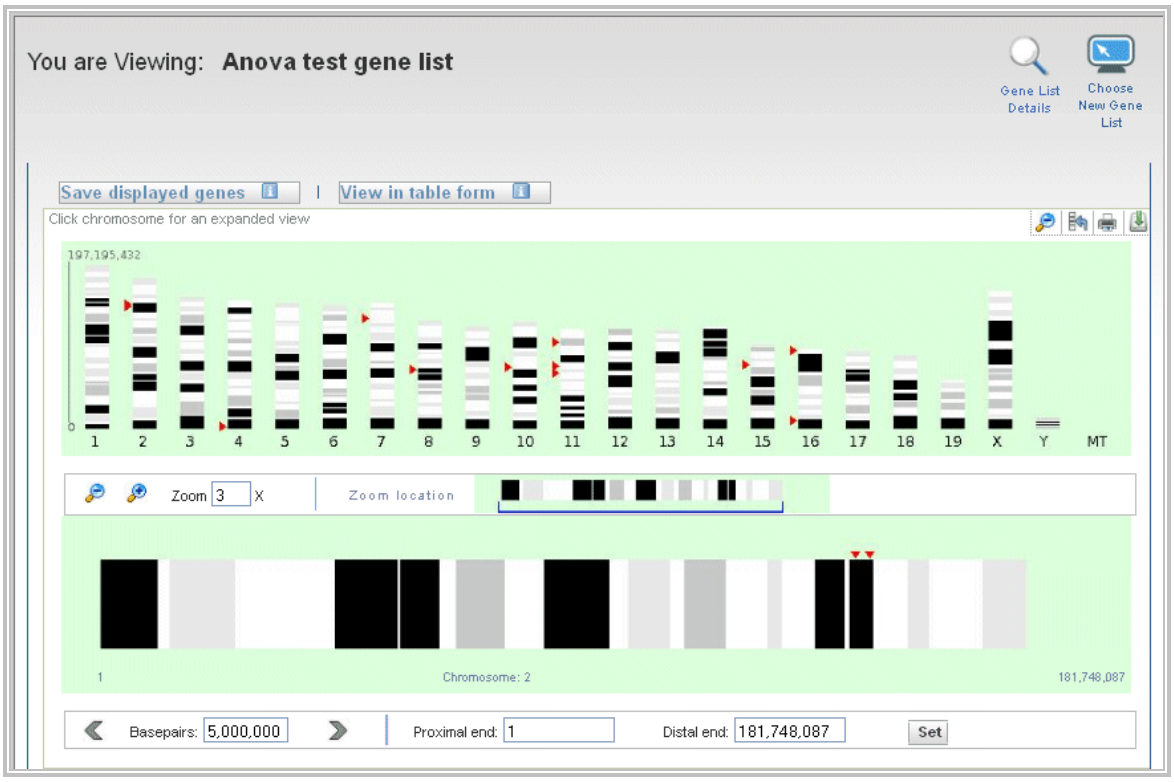
- Click the **View in table form** link to see a table of all of the genes currently displayed on the graphic. If any of the probesets for a given gene have a significant eQTL, all of the probesets for that gene are displayed in the table, but only the most significant eQTL position ("Max LOD") and the associated marker locations are shown on the graphic.
    - Click the **View** link in the gene table to display individual LOD plots for each probeset. Copy these by right-clicking on the image.
    - Click the **Download** button to download the table into a tab-delimited text file.
    - Click the **Customize this view** button to change the default view of the table, which includes all probesets for a given gene.
        - Choose **Genes (and all associated probesets) from the [selected] list that meet the restriction criteria** to restrict the list to those with eQTLs (probesets) that overlap the selected bQTL intervals.
        - Select **Probesets that are in the [selected] gene list** to restrict the list to probesets (as opposed to genes) whose expression values are significantly correlated with the phenotype. These probesets are also identified by the asterisk in the first column of the table.
        - Selelct **Probesets that have eQTL p-value < [selected value]** to restrict the list to only the probesets passing the desired threshold.
        - Choose **Genes (and all associated probesets) from the [selected] list with probesets that did not meet the restriction criteria or were not considered in eQTL** to restrict the list to those that were not considered in eQTL.
- Click the **Download** icon to download the chromosome map as a JPEG graphic.
- Click the **Save displayed genes** link to save the displayed genes as a gene list.

- Click the **Magnify** icon in the graphic to expand the graphic to the full page.
- Click a chromosome in the graphic to show an expanded view. See "Expanded Chromosomal View".
- Set the *Gene Graphic Settings*:
  - **Show physical location of genes** - Choose this option to mark the physical location of genes with a red arrow.
  - **Show genomic locations of transcription control in...**Choose this option to mark the transcription control locations with a green arrow, and choose a corresponding eQTL p-value. For mouse, only brain is available. For rat, options are brain, heart, liver, and brown adipose.
- Click **Advanced Settings** to enter user-defined regions, such as QTLs. You can also:
  - Click **View** to see details about the specific region.
  - Click the **Delete** icon beside the region you want to delete.



Expanded Chromosomal View

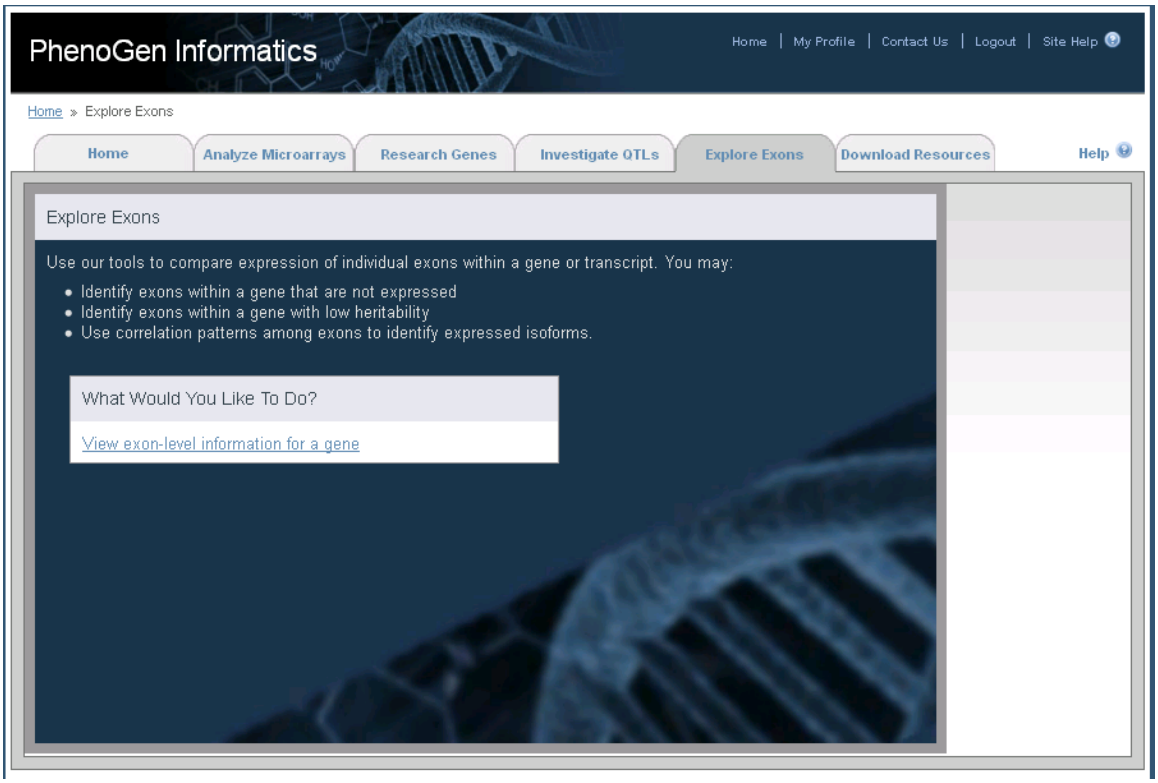
The expanded chromosomal view zooms in on the specific location you choose in the graphic, shows the zoom location, and allows you to set basepair start (left) and end (right) positions.


You are Viewing: Anova test gene list
Gene List Details
Choose New Gene List
Save displayed genes
View in table form
Click chromosome for an expanded view
197,195,432
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y MT
Zoom 3 X
Zoom location
1
Chromosome: 2
181,748,087
Basepairs: 5,000,000
Proximal end: 1
Distal end: 181,748,087
Set

Exploring Exons

The **Explore Exons** tab allows you to view the correlation among exons within a transcript, using expression data from two recombinant inbred panels of rodents in our public data sets:

- HXB/BXH for rats
- ILSXISS for mice



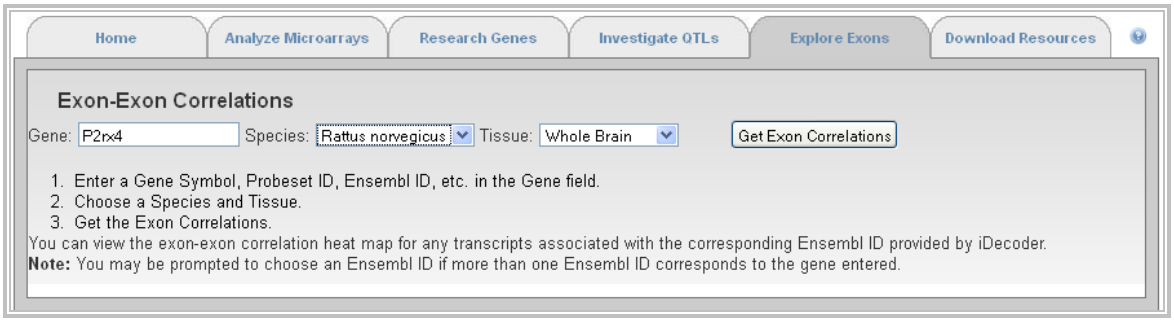
Choose **View exon-level information for a gene** to get started. See "Viewing Exon-level Correlations" for details.

Viewing Exon-level Correlations

You can create an exon correlation heatmap for a specific gene, species, and tissue. The data is retrieved from multiple databases, so generating the initial graphics may take a few moments.You can display the heat-maps for two transcripts of a gene side-by-side to determine the transcript that best fits the expression cor-relation patterns.

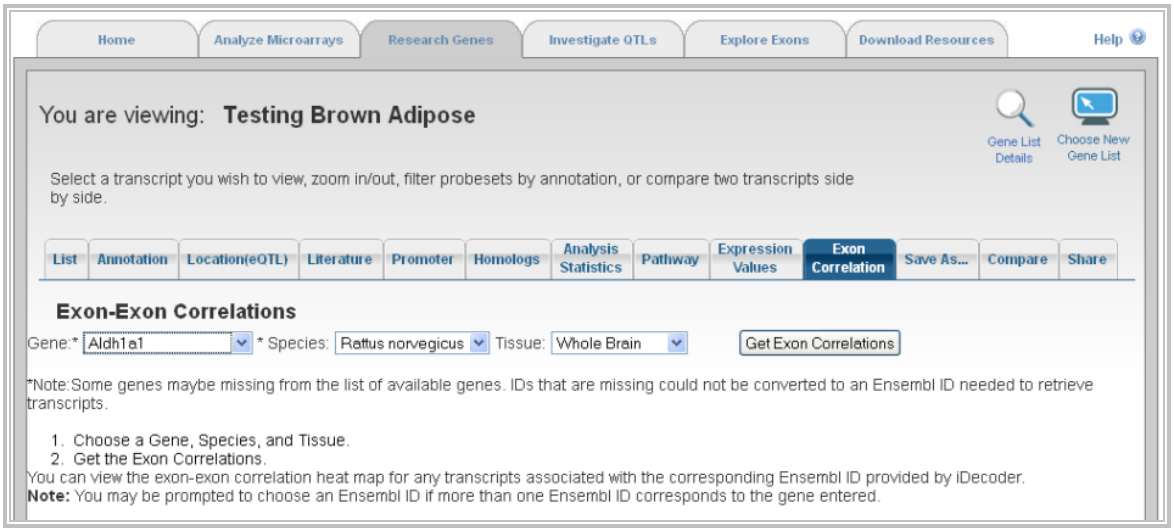
Explore Exons tab:

1. Click the **Explore Exons** tab. The *Explore Exons* tab displays.
2. Click **View exon-level information for a gene** in the *What would you like to do* section. The *Exon-Exon Correlations* page displays.



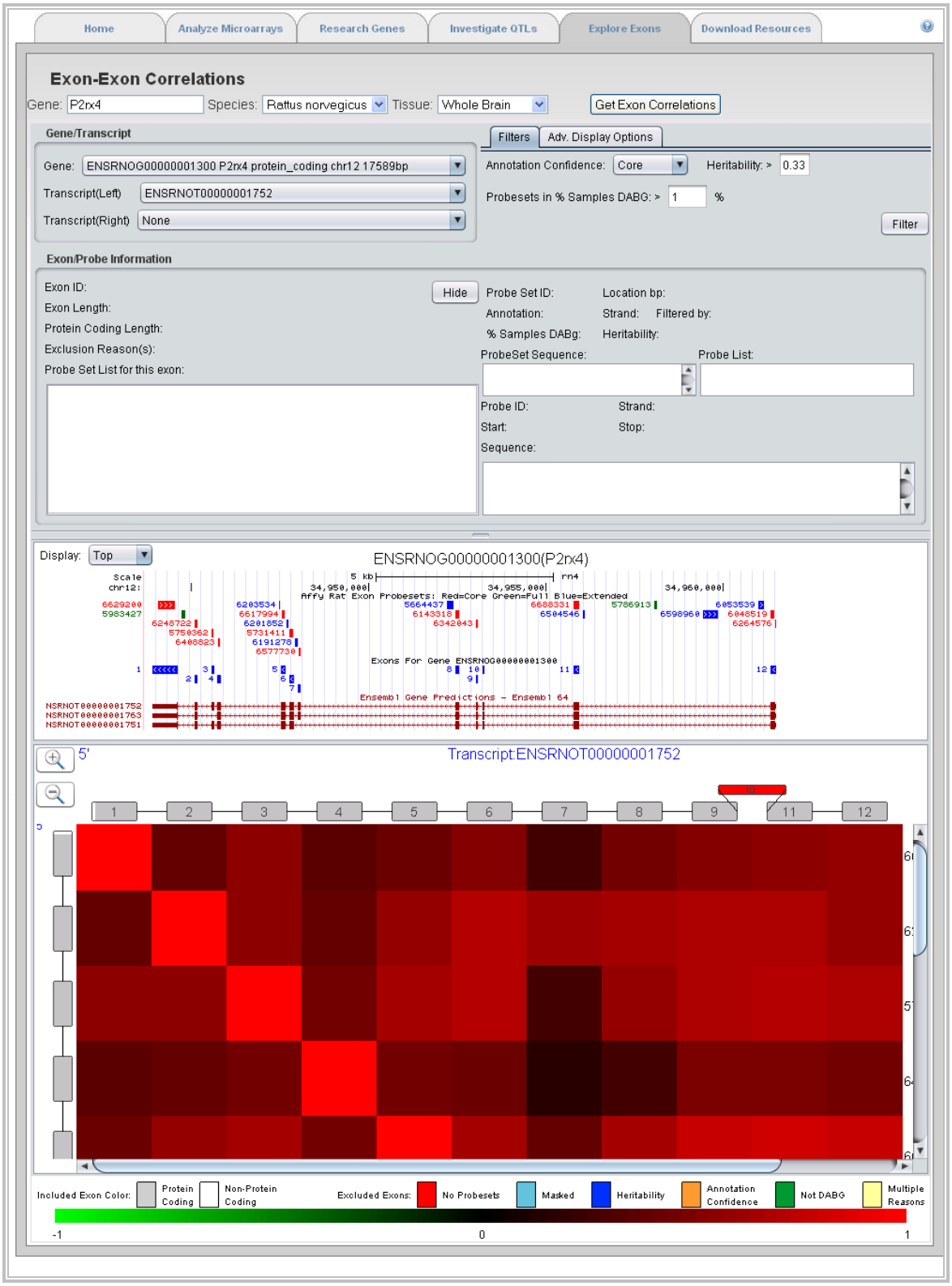
Research Genes tab:

1. Click the **Research Genes** tab. The *Research Genes* page displays.
2. Click **Select a gene list for further investigation** in the *What would you like to do* section. A page displays the gene lists to which you have access.
3. Click the gene list for which you want to view gene expression data.
4. Click the **Exon Correlation**tab. The *Exon-Exon Correlations* page displays.



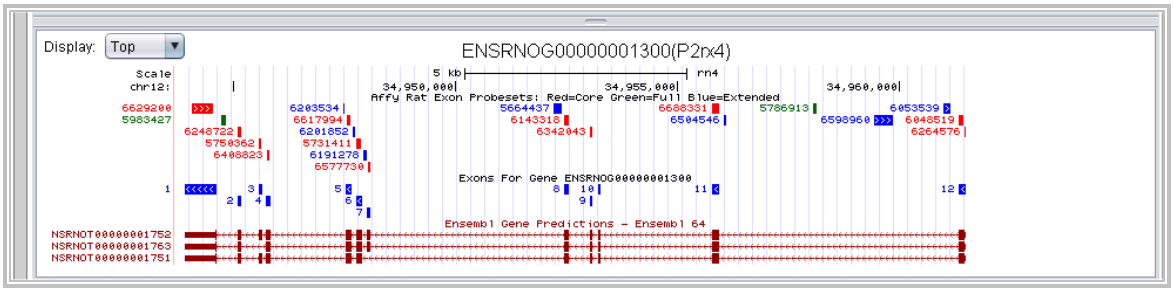
On the Exon-Exon Correlations page:

1. Enter the **Gene** for which you want to create an exon correlation. On the *Exon Correlation* tab, the available genes for the selected Gene List are available in a drop-down list.
2. Select the **Species** and the **Tissue** to which this correlation pertains.
3. Click **Get Exon Correlations**. The exon correlations display.



4. **OPTIONAL:** Display two heat maps by selecting both a Left and Right Transcript in the In the *Gene/Transcript* section at the top left.
5. **OPTIONAL:** Hover over the schematic on top of the heatmap or click on individual exons in the schematic to find out more information about individual exons and the probesets that align to them.

By default, the first graphic, which is generated through the University of California Santa Cruz Genome Browser (http://genome.ucsc.edu/), displays the probesets from the exon array and the annotated Ensembl transcripts (i.e., isoforms) for the input gene.



The exons are color-coded based on their level of annotation according to Affymetrix. The red probesets are included in the core set of probesets and indicate high confidence in the annotation. The blue probesets are included in the extended set of probesets according to Affymetrix, and the green probesets are from the full set of probesets and indicate the least amount of confidence in the annotation. The next track assigns a number to each annotated exon of the gene to aid in interpretation of the heatmap that follows.

By default, the second graphic is a heatmap that displays the correlations among exons of a given transcript of the gene.

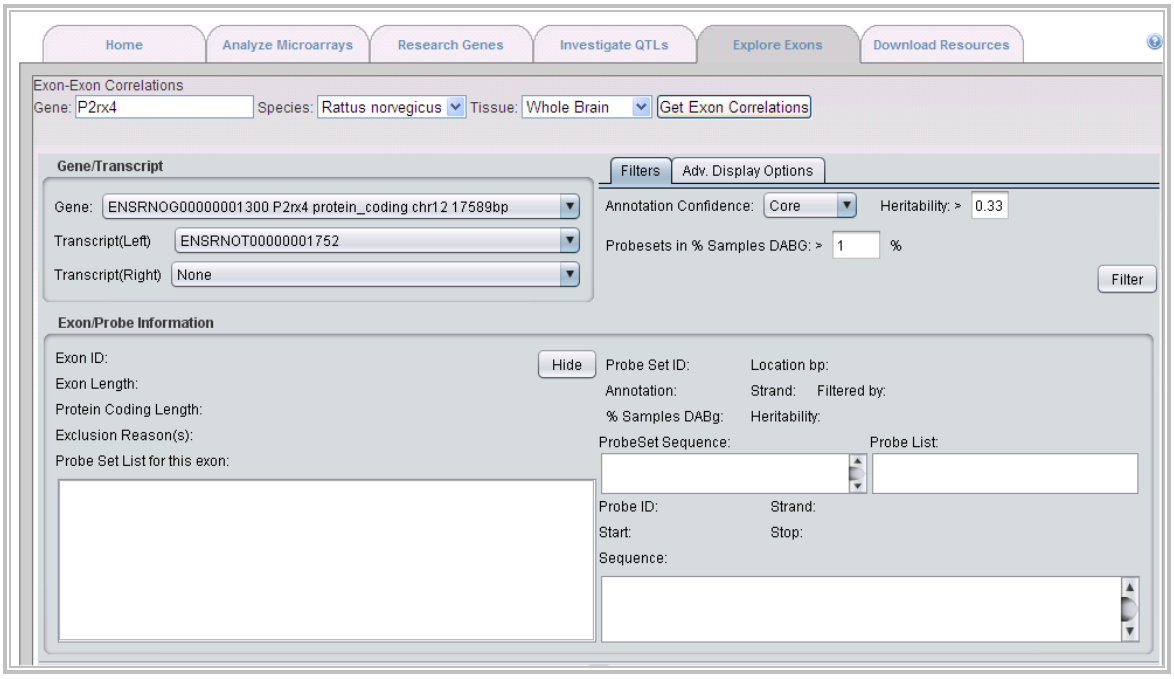
Each row and column represents a probeset that passed the filtering criteria specified at the top right side of the page. Along the top of the heatmap is a schematic of the exon structure of the transcript. The boxes represent exons of the transcript and are numbered, as in the top figure. Grey boxes are exons that are probed by the array, met the filtering criteria, and are displayed in the heatmap. Exons that do not satisfy these criteria are colored based on the reason why they are not included in the heatmap.

- **No Probesets (Red)** indicates that the Affymetrix array does not contain any probesets that target this exon.
- **Masked (Light Blue)** indicates that the probeset was removed due to poor intergrity of its probes (i.e., the probes aligned to many regions of the genome or targeted a region of the genome that harbored a known SNP).
- **Heritability (Blue)** indicates that the probeset had a low (according to filtering criteria) heritability in the RI panel.

- **Annotation Confidence (Orange)** indicates that probesets that align to this exon are not included due the annotation confidence criteria selected for filtering.
- **Not DABG (Green)** indicates that the probesets that align to this exon are not expressed above background according to the DABG criteria selected for filtering.
- **Multiple Reasons (Yellow)** indicates that the exon was filtered out for more that one reason, e.g., low heritability and not DABG

In the top right section, choose the probesets to display for the heatmap based on the:

- Annotation Confidence from Affymetrix (Core, Extended, or Full).
- Heritability of the probeset's expression in the RI panel.
- Proportion of samples with expression values above background (Detected Above Background–DABG).

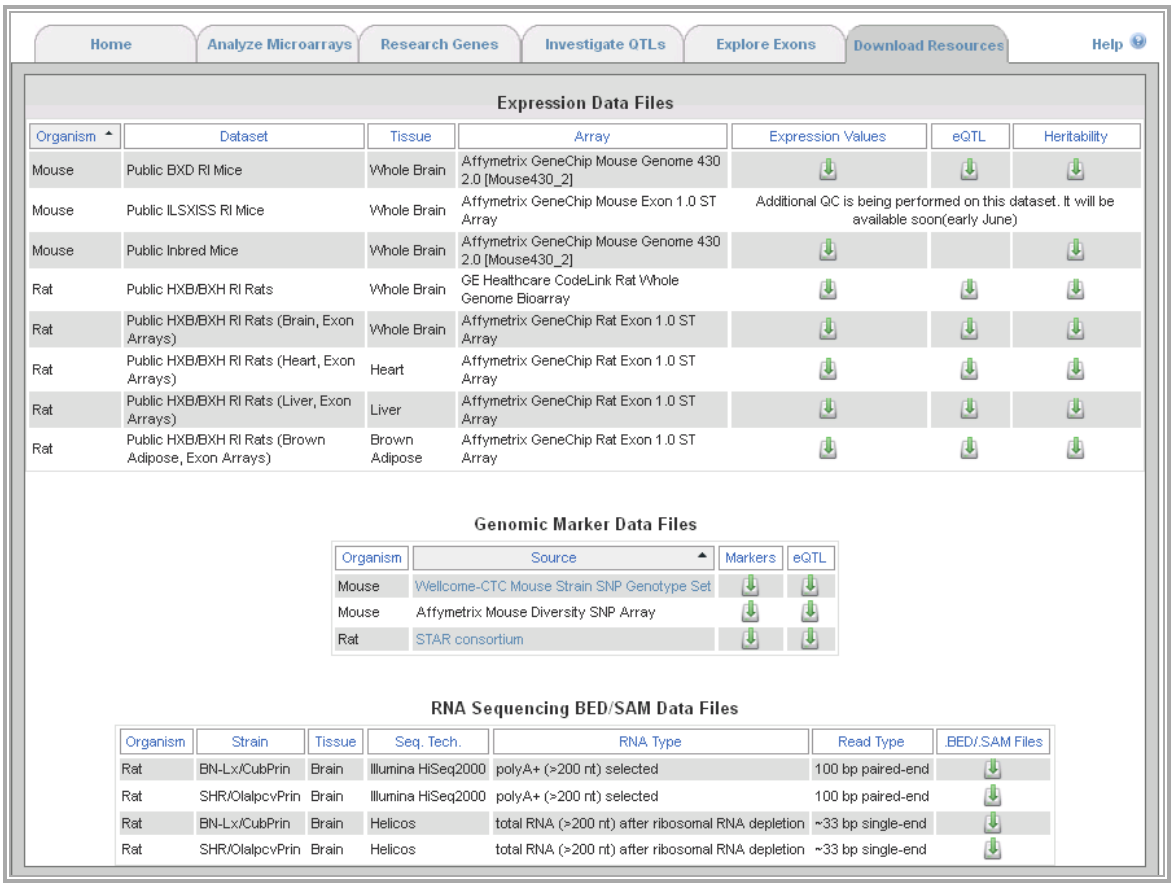


There are also advanced display options that allow you to include probesets that align to an intronic region of the gene or align to the opposite strand from which the gene is coded.

Download Resources

The **Download Resources** tab allows you to download:

- Expression Data Files (Expression Values, eQTL, and Heritability).
- Genomic Marker Data Files (Markers and eQTL).
- RNA Sequencing BED/SAM Data Files

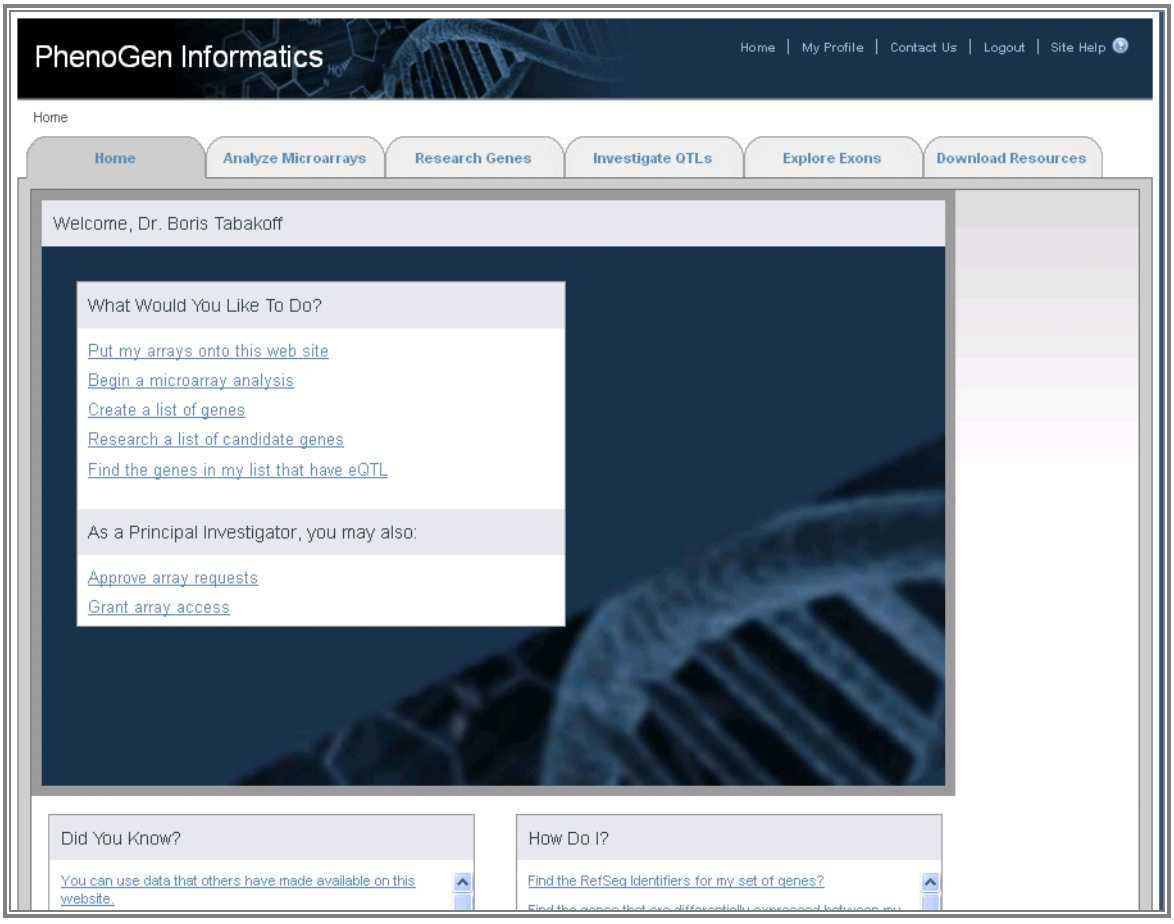


Principal Investigator Overview

Only users with **Principal Investigator permissions** can see the Principal Investigator pages.

The Principal Investigator (PI) is often the head of a lab, is responsible for the array data, and must grant permission to other users to view arrays which s/he is responsible for. When arrays are uploaded, the PI can assign permission individually to any users that s/he wants to have access to the data. If a user wishes to gain access to an array, the user can add the array to a dataset, and a request for permission is sent to the PI who owns the array. The PI can also make array data accessible to all.

A user who is a Principal Investigator sees a Principal Investigator menu on the *Home* tab after Login, below the *What would you like to do* box.



Choose an option to get started:

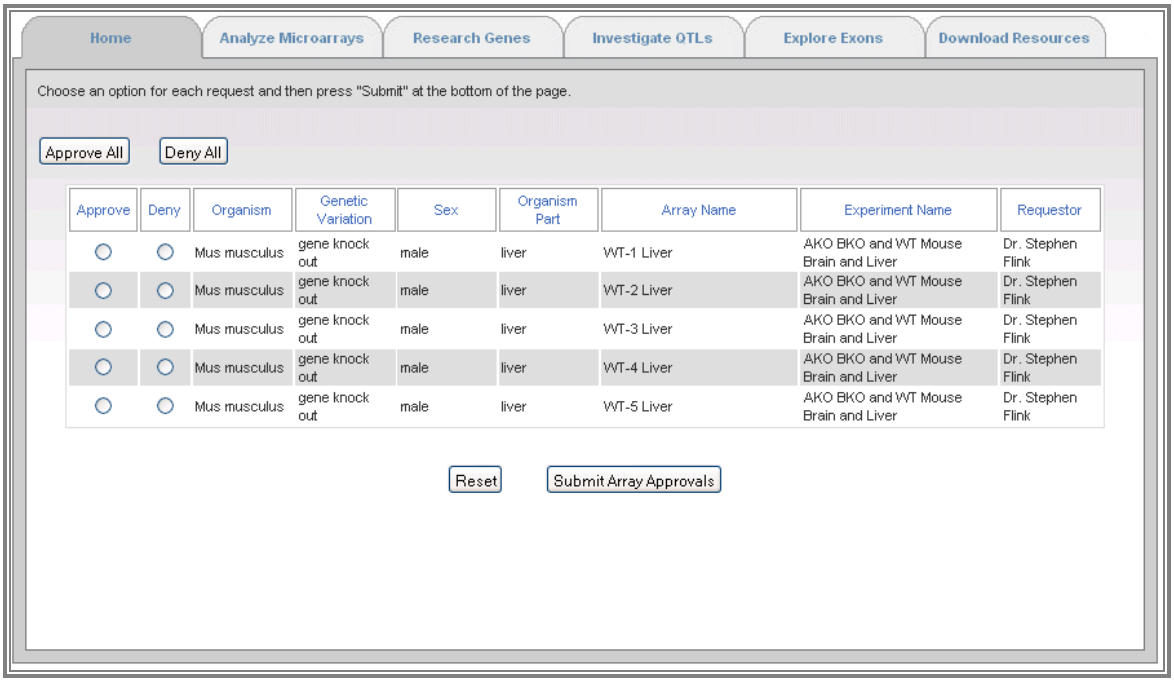
- Approve array requests. See "Approving Array Requests".
- Grant array access to an individual. See "Granting Array Access".
- Grant open access to a set of arrays. See "Granting Array Access".

Approving Array Requests

Only users with **Principal Investigator permissions** can approve pending requests for arrays.

The *Approval* page allows a Principal Investigator to review the users who have requested access to his arrays. The PI can approve or deny access to the array.

1. Click the **Home** tab. The *Home* page displays.
2. Click **Approve array requests** in the *Principal Investigator* section below the *What would you like to do* section. The *Approval* page displays.
3. Click **Approve** or **Deny** next to each request, or click **Approve All** or **Deny All**.
4. Click **Submit Array Approvals**. An email is sent to each requester that informs them whether their array request(s) has been approved or denied.



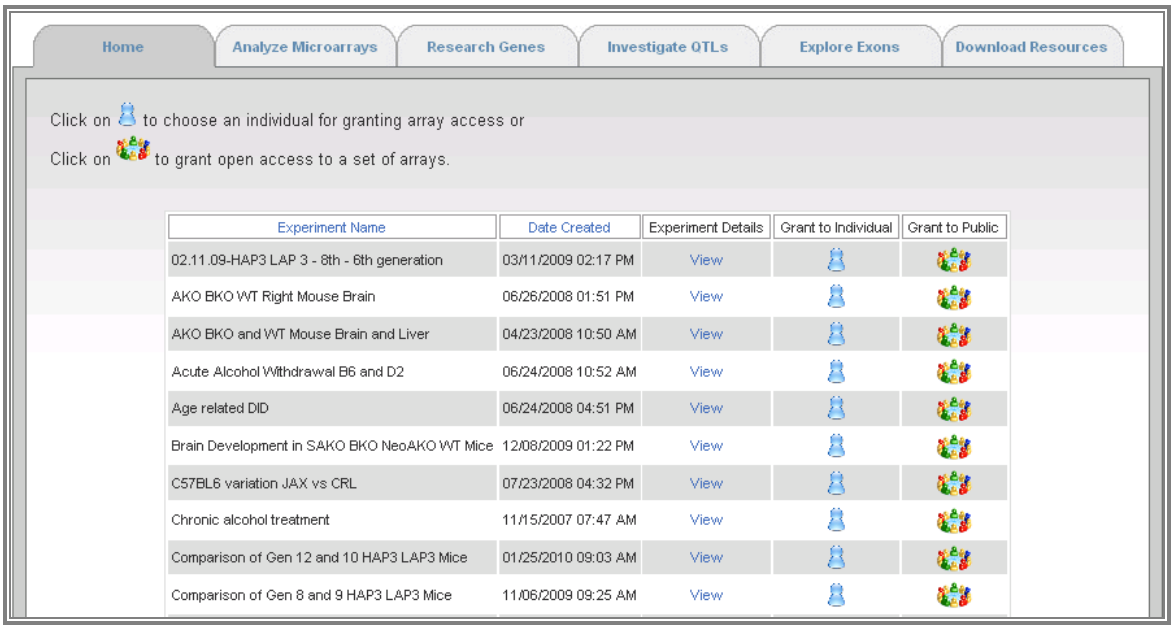
Granting Array Access

Only users with **Principal Investigator permissions** can grant array access.

The *Grant Access* page allows a Principal Investigator to give an individual user access to the arrays for which s/he is responsible. It also allows a PI to give open access to the array data in a MIAME-compliant experiment for the use of ***ALL*** registered users of the PhenoGen website.

Granting Access

1. Click the **Home** tab. The *Home* page displays.
2. Click **Grant array access** in the *Principal Investigator* section below the *What would you like to do* section. The *Grant Access* page displays.

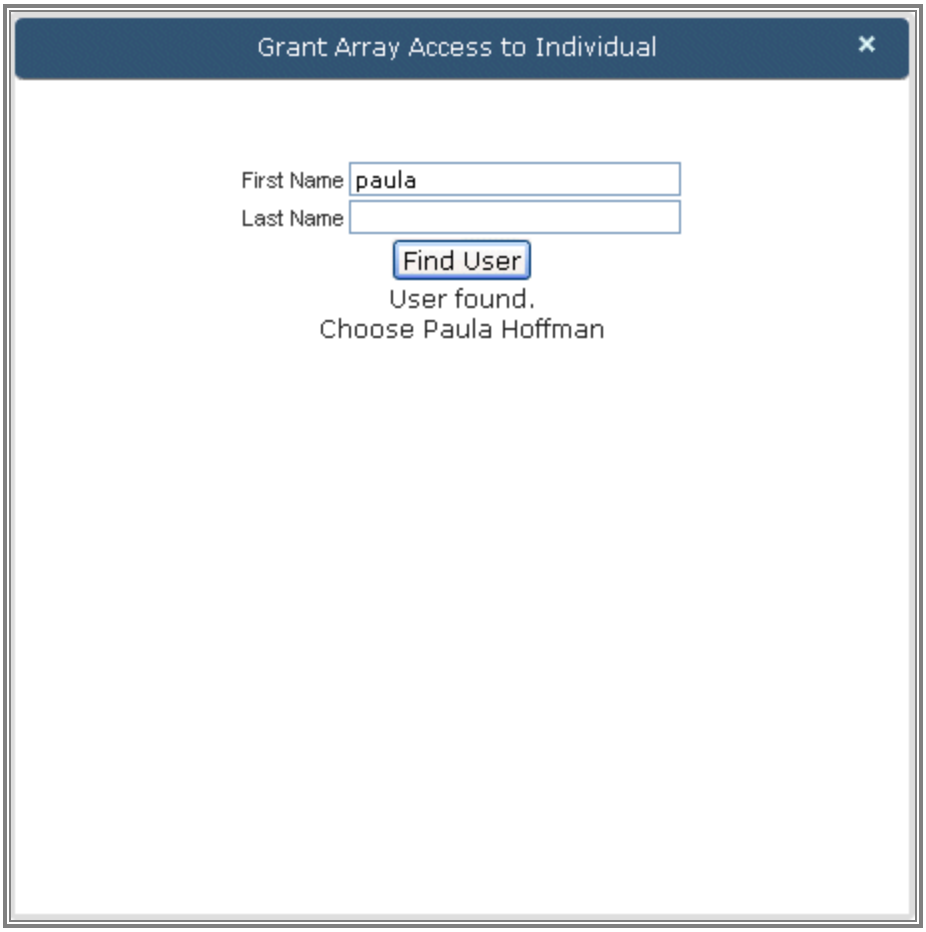


Note: Click **View** in the Experiment Details column to see array details for an experiment.

For an Individual

On the Grant Access page:

1. Click the **Grant to Individual**icon in the row of the experiment to which you want to grant access.
2. Enter the first or last (or both) names of the individual.
3. Click **Find User**. Any matches display.



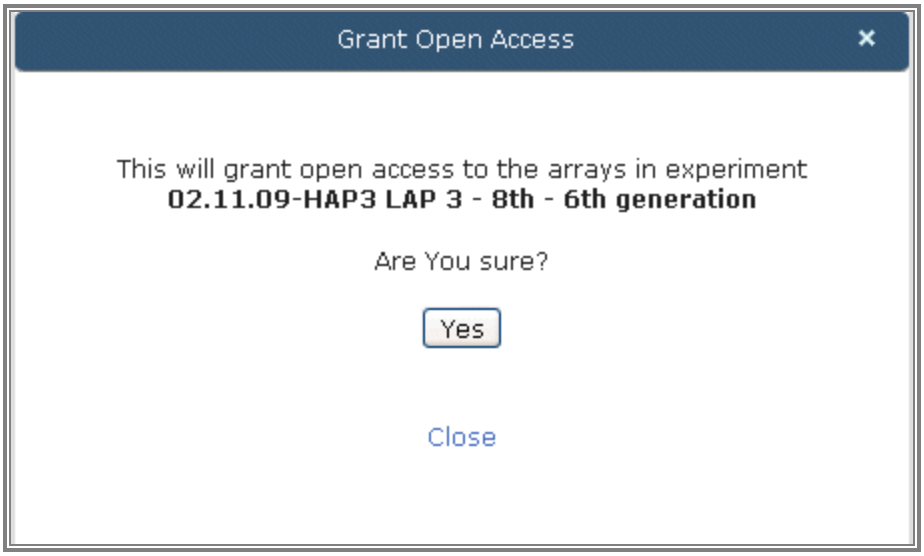
4. Click **Choose [User Name]** to give that user access to the array.

Note: This only gives access to the current arrays in the PhenoGen array database. Access must be granted as experiments are added.

For Everyone

On the Grant Access page:

1. Click the **Grant to Public** icon in the row of the experiment to which you want to grant access.
2. Click **Yes**. The data in the chosen experiment is now available to all registered users of the PhenoGen website.



Supplementary Information

Supplementary Information contains additional details about various topics. See:

Additional Quality Control Sources

Some of the QC procedures for commonly used microarrays can be found on the manufacturer's websites. For further details, visit the following URLs:

- **Affymetrix:** http://www.affymetrix.com
- **CodeLink:** http://www.codelinkbioarrays.com

Additional computational QC tools for the R Statistical Package can be found on:

http://bioinformatics.picr.man.ac.uk/research/software/simpleaffy/qcstats.html

http://plmimagegallery.bmbolstad.com/

All About R

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

Citation for R

R Development Core Team (2006). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Citation for Bioconductor

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5:R80.

For more details about R, see the topics *Viewing the R Project Homepage* and *Viewing R Manuals*.

Viewing the R Project Homepage

1. Open a new browser window.
2. Enter the URL http://lib.stat.cmu.edu/R/CRAN/
3. Click **R Homepage** in the vertical menu on the left. The *R Homepage* displays.

Viewing R Manuals

1. Open a new browser window.
2. Enter the URL http://lib.stat.cmu.edu/R/CRAN/

3. Click **Manuals** in the vertical menu on the left. The *R Manuals* page displays.
4. Select and click the manual you want to view.

Expression QTL Derivation

For both mice and rats, mean expression levels within strains were used as phenotypic values in a QTL analysis implemented in QTL Reaper, which is written in C and compiled as a Python module. QTL Reaper can be downloaded from http://qtlreaper.sourceforge.net/. A weighted marker regression analysis was used within QTL Reaper to calculate LRS scores for each marker. LRS scores were transformed to LOD scores for convenience by dividing by (2 x ln(10)). The regression is weighted to account for the different number of arrays within strains used to calculate strain means. The weight is based on the repeatability of the transcript intensity and the number of arrays used to calculate the strain mean (Carlborg et al. 2005). The empirical p-value with respect to the maximum LOD score was calculated for each transcript by permutation (Churchill and Doerge, 1994). The empirical p-value adjusts for the multiple comparisons due to the multiple markers per transcript, but not for the multiple comparisons due to the many transcripts. To adjust for the multiple comparisons due to many transcripts, false discovery rates were calculated according to Benjamini and Hochberg (1995).

Mouse, whole brain, Affymetrix Mouse 430 version 2 array

Whole-brain gene expression data was obtained for a panel of 30 BXD RI strains plus the two parental strains on the MOE430v2 array from Affymetrix. Probes were eliminated prior to normalization if their sequence did not match any part of the NCBI m37 build of the mouse genome, if their sequence matched multiple locations in the mouse genome, or if the location in the genome that the probe did match contain a single nucleotide polymorphism between C57BL/6J and DBA/2J according to the whole genome sequence data obtained from the Sanger Institute (Keane et al. 2011). Entire probe sets were eliminated if less than four of the original probes remained after filtering. Probe set intensities were normalized and summarized using RMA. If a probeset did not have at least one present call throughout all samples, the probeset was dropped from the data set. Of the 41,581 probesets retained after masking, 30,031 probesets remained after filtering by present/absent calls. Data were thoroughly examined for batch effects related to processing. The microarrays were run over a year and a half period, resulting in 15 batches. Both batches and strains can contribute to non-random data distribution and a new method for removing batch effects, while retaining strain effects, was used (personal communication, Evan Johnson, Boston University) on the set of 30,031 probesets detected above background. This method combines a simple rank test and a Bayesian hierarchical framework similar to the empirical Bayes method, Combating Batch Effects When Combining Batches of Gene Expression Microarray Data (ComBat) (Johnson et al., 2007). This version of the data is available in the Download Resources section. The two parental strains were not included in eQTL calculations.

An original set of 115,183 SNPs markers and their genotypes for 89 BXD RI strains and the two parental strains was downloaded from the Wellcome-CTC Mouse Strain SNP Genotype Set (http://gscan.well.ox.ac.uk/gsBleadingEdge/mouse.snp.selector.cgi). The location of these markers is based on Mouse Build 37/mm9. However, the set of markers used for the eQTL analyses was reduced from the original set to eliminate SNPs with missing genotype information for the 30 RI strains, SNPs that did not differ between the RI strains, and SNPs with genotype calls that did not follow the known breeding scheme of the panel. This reduced the SNP set to 7,226 SNPs. This set of SNPs was reduced to unique strain distribution patterns with respect to the 30 RI strains used in our analysis. This final set contained 986 informative strain distribution patterns. Both the normalized expression data and the markers used for the eQTL analysis are available for download from the PhenoGen website.

Mouse, whole brain, Affymetrix Mouse Exon Array

Whole-brain gene expression data was obtained for a panel of 59 LXS RI strains on the Affymextrix Mouse Exon Array 1.0 ST. Individual probes were eliminated prior to normalization if their sequence did not match any part of the NCBI m37 build of the mouse genome, if their sequence matched multiple locations in the mouse genome, or if the location in the genome that the probe did match contained a SNP between any of the 19 strains in the public Inbred Mice dataset where genotype data is available at the Imputed Genotype

Resource from the Jackson Laboratory; http://cgd.jax.org/datasets/popgen/imputed.shtml (same mask that is implemented on PhenoGen). Entire probe sets were eliminated if less than three of the original probes remained after filtering.

Arrays were examined for quality and arrays that did not meet quality standards were eliminated. Probe intensities were normalized and summarized into core transcript clusters using RMA. The dataset, including arrays from ILS, ISS, C57BL/6J, and DBA/2J, was adjusted for batch effects using the empirical Bayes method outlined in Johnson et al (2007). Two C57BL/6J arrays were analyzed in each batch, and most batches also had two DBA/2J arrays. Strain means were calculated after adjusting for the effect of breeding location. For most strains, three animals were bred at the Jackson Laboratory in Bar Harbor, Maine and three animals were bred at the University of Colorado School of Medicine in Aurora, Colorado. The marker set of eQTL calculations used on the LXS RI panel came from Gary Churchill at the Jackson Laboratory. SNP genotypes were assessed using the Affymetrix Mouse Diversity Genotyping Array. Of the 314,865 SNPs retrieved, 34,475 SNPs indicated different homozygous genotypes between parental strains (ILS and ISS), had valid dbSNP identifiers, and had no missing or heterozygous genotype calls. The set of markers used for the eQTL analyses was reduced from the original set to eliminate markers that had identical strain distributions (with respect to the 59 strains used in our analysis). This final set contained 1,475 informative markers. Both the normalized expression data and the markers used for the eQTL analysis are available for download from the PhenoGen website.

Rat, whole brain, CodeLink Whole Genome Rat Array

Whole-brain gene expression data was obtained for a panel of 25 HXB/BXH RI strains plus the two parental strains on the CodeLink Rat Whole Genome Array. In preparation for normalization, probes were removed from the datasets if they were one of the negative or positive controls placed on the array by the manufacturer. Next, individual values were eliminated based on the quality flags assigned by the CodeLink Expression Analysis Software. Values were eliminated if they were flagged as M (spot was identified to be defective through image inspection at manufacturing), C (spot has a high level of background contamination), I (spot has an irregular shape), or S (spot has a high number of saturated pixels). Values were retained if they were flagged G (spot is good) or L (spot is below local background noise). In addition, to be able to take the log base 2 transformation of the background adjusted intensity values, all background adjusted intensity values below zero were replaced with the value 0.00001. The data were then normalized using a cyclic LOESS procedure executed in R, which accounted for the missing intensity values. Genotype information for the rats was downloaded from the STAR Consortium's website (http://www.snp-star.eu/). SNP locations are based on RGSC version 3.2. The downloaded SNP data was cleaned by eliminated SNPs that did not differ between the parental strains, SNPs that are not genotyped in either parental strains, and SNPs that were heterozygous for either of the parental strains. Unknowns were recoded if the surrounding SNPs had the same genotype. Double recombinants were also recoded. SNPs were eliminated if more than two strains were missing genotype information. After the dataset had been cleaned, 1,460 unique strain distribution patterns were identified and used in the eQTL analysis. Both the normalized expression data and the markers used for the eQTL analysis are available for download from the PhenoGen website.

Rat, whole brain/left ventricle/liver/brown adipose tissue, Affymetrix Rat Exon Array

Whole-brain, heart, liver, and brown adipose tissue gene expression data was obtained for a panel of 21 HXB/BXH RI strains (only 19 RI strains for the brown adipose tissue) and six related inbred strains on the Affymextrix Rat Exon Array 1.0 ST. Individual probes were eliminated prior to normalization if their sequence did not match any part of the RGSC version 3.2 of the rat genome, if their sequence matched multiple locations in the mouse genome, or if the location in the genome that the probe did match contain a SNP between the Brown Norway (BN/SsNHsdMcwi) inbred strains (reference strain) and the spontaneously hypertensive rat (SHR/OlaIpcv) strain that was recently sequenced (Atanur et al. 2010) using next generation sequencing or a SNP detected in DNA sequencing of the BN-Lx/CubPrin and SHR/OlaIpcvPrin strains (same mask that

is implemented on PhenoGen). DNA sequence data for the BN/SsNHsdMcwi and SHR/OlaIpcv was downloaded directly from the Ensembl ftp site at: ftp://ftp.ebi.ac.uk/pub/databases/ensembl/snp/rat/shr/. For the 4,022,111 original probes, 604,601 were removed (472,072 did not map uniquely to the genome; 132,529 probes contained a SNP). Entire probe sets were eliminated if less than three of the original probes remained after filtering. Arrays were examined for quality, and arrays that did not meet quality standards were eliminated. Probe intensities were normalized and summarized into core transcript clusters using RMA. The dataset, including arrays from the six relevant inbred strains, was adjusted for batch effects using the empirical Bayes method outlined in Johnson et al (2007). Only the 21 recombinant inbred strains (19 for the brown adipose tissue data set) were included in the eQTL analysis. The marker set used of eQTL calculations on the HXB/BXH RI panel was originally downloaded from the Ensembl link to the STAR Consortium data (http://www.ensembl.org/Rattus_norvegicus/Info/Content?file=star/index.html). SNP locations are based on RGSC version 3.2. The downloaded SNP data was cleaned by eliminated SNPs that did not differ between the parental strains, SNPs that are not genotyped in either parental strains, and SNPs that were heterozygous for either of the parental strains. Unknowns were recoded if the surrounding SNPs had the same genotype. Double recombinants were also recoded. SNPs were eliminated if more than two strains were missing genotype information. After the dataset had been cleaned, 761 unique strain distribution patterns were identified and used in the eQTL analysis. Both the normalized expression data and the markers used for the eQTL analysis are available for download from the PhenoGen website

References

1. Johnson WE, Li C, and Rabinovic A (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1): 118-127.
2. Atanur SS, Birol I, Guryev V, Hirst M, Hummel O, Morrissey C, Behmoaras J, Fernandez-Suarez XM, Johnson MD, McLaren WM, Patone G, Petretto E, Plessy C, Rockland KS, Rockland C, Saar K, Zhao Y, Carninci P, Flicek P, Kurtz T, Cuppen E, Pravenec M, Hubner N, Jones SJ, Birney E, Aitman TJ (2010). The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Research 20(6):791-803.
3. Churchill GA and Doerge RW (1994). Empirical threshold values for quantitative trait mapping. Genetics 138:963-971.
4. Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57:289-300.

MIAME Overview

MIAME (Minimum Information About a Microarray Experiment) is a standard for exchanging microarray experimental data in such a way as to make it easily interpreted and allow for easy and independent verification. It is described in the abstract of the original MIAME proposal [1] as follows:

> Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

Thus MIAME compliance provides the minimum information required to interpret unambiguously and potentially reproduce and verify an array-based gene expression monitoring experiment. Although details for particular experiments may be different, MIAME aims to define the core that is common to most experiments. MIAME is not a formal specification, but a set of guidelines.

One of the major objectives of MIAME is to guide the development of microarray databases and data management software. A standard microarray data model and exchange format MAGE [2], which is able to capture information specified by MIAME, has been submitted by EBI (for MGED) and Rosetta Biosoftware, and recently became an Adopted Specification of the OMG standards group. Many organizations, including Agilent, Affymetrix, and Iobion, have contributed ideas to MAGE.

Although MIAME concentrates on the content of the information and should not be confused with a data format, it also tries to provide a conceptual structure for microarray experiment descriptions [3].

It is therefore of crucial importance that all users of the PhenoGen site closely conform to these guidelines. To ensure this, the website is structured in such a way that very little room is given for non-conformity.

References

[1] Nature Genetics, December 2001, 29:365 – 371. http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v29/n4/abs/ng1201-365.html

[2] http://www.mged.org/Workgroups/MAGE/mage.html

[3] http://www.mged.org/Workgroups/MIAME/miame_1.1.html

Promoter Analysis Tools

Program	Operating Principle	Technical Data and URL	Reference
AlignACE	Gibbs sampling algorithm that returns a series of motifs as weight matrices that are over-represented in the input set.	Judges alignments sampled during the course of the algorithm using a maximum a priori likelihood score, which gauges the degree of over-representation. Provides an adjunct measure (group specificity score) that takes into account the sequence of the entire genome and highlights those motifs found preferentially in association with the genes under consideration. http://atlas.med.harvard.edu	1
ANN-Spec	Models the DNA binding specificity of a transcription factor using a weight matrix.	Objective function based on log likelihood that a transcription factor binds at least once in each sequence of the positive training data compared with the number of times it is estimated to bind in the background training data. Parameter fitting is accomplished with a gradient descent method, which includes Gibbs sampling of the positive training examples. http://www.cbs.dtu.dk/services/DNAarray/ann-spec.php	2
Consensus	Models motifs using weight matrices searching for lhe matrix with maximum information content.	Uses a greedy method, first finding the pair of sequences that share the motif with greatest information content, then finding the third sequence that can be added to the motif, resulting in greatest information content. http://bifrost.wustl.edu/consensus/	3
GLAM	Gibbs sampling-based algorithm that automatically optimizes the alignment width and evaluates the statistical significance of its output.	Since the basic algorithm cannot find multiple motif instances per sequence, long sequences are fragmented into shorter ones, and the alignment is transformed into a weight matrix and used to scan the sequences to obtain the final site predictions. http://zlab.bu.edu/glam/	4
Improbizer	Uses expectation maximization to determine weight matrices of DNA motifs that occur improbably often in the input sequences.	As a background (null) model it uses up to a second-order Markov model of background sequence. Optionally, Improbizer constructs a Gaussian model of motif placement so that motifs that occur in similar positions in the input sequences are more likely to be found. http://www.soe.ucsc.edu/~kent/improbizer	5
MEME	Optimizes the E-value of a statistic related to the information content of the	Rather than sum of information content of each motif column, the statistic used is the product of the p-values of column information contents. The motif search consists of performing expectation maximization from starting points derived from each subsequence occurring in the input sequences. MEME differs from MEME3 mainly in using a correction factor to improve the accuracy of the objective function.	6

	motif.	http://meme.sdsc.edu	
MITRA	Uses an efficient data structure to traverse the space of IUPAC patterns.	For each pattern, MITRA computes the hypergeometric score of the occurrences in the target sequences relative to the background sequence and reports the highest scoring patterns. http://www.ccls.columbia.edu/compbio/mitra/	7
MotifSampler	Matrix-based motif finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model.	The probabilistic framework is further exploited to estimate the expected number of motif instances in the sequence. http://www.esat.kuleuven.ac.be/~dna/Biol/Software.html	8
Oligo/dyad analysis	Detects over-represented oligonucleotides with oligo analysis and spaced motifs with dyad analysis.	These algorithms detect statistically significant motifs by counting the number of occurrences of each word or dyad and comparing these with expectation. The most crucial parameter is the choice of an appropriate probabilistic model for the estimation of occurrence significance. http://rsat.ulb.ac.be/rsat/	9, 10
QuickScore	Based on an exhaustive searching algorithm that estimates probabilities of rare or frequent words in genomic texts.	Incorporates an extended consensus method allowing well-defined mismatches and uses mathematical expressions for efficiently computing z-scores and p-values depending on the statistical models used in their range of applicability. Special attention is paid to the drawbacks of numerical instability. The background model is Markovian, with order up to 3. http://algo.inria.fr/dolley/QuickScore/	11
SeSiMCMC	Modification of Gibbs sampler algorithm that models the motif as a weight matrix, optionally with the symmetry of a palindrome or of a direct repeat and optionally, with spacers.	Includes two alternating stages. The first one optimizes the weight matrix for a given motif and spacer length. The algorithm changes the positions of the motif occurrence in the sequences and infers the motif model from the current occurrences. These changes are used to optimize the likelihood of sequences as being segmented into the (Bernoulli) background and the motif occurrences. The optimization is organized via a Gibbs-like Markov chain that samples positions in sequences one-by-one until the Markov chain converges. The second stage looks for best motif and spacer lengths for obtained motif positions. It optimizes the common information content of motif and of distributions of motif occurrence positions. http://favorov.imb.ac.ru/SeSiMCMC/	12
Weeder	Consensus-based method that enumerates exhaustively all the oligos up to maximum length and collects	Each motif is evaluated according to the number of sequences in which it appears and how well conserved it is in each sequence, with respect to expected values derived from the oligo frequency analysis of all the available upstream sequences of the same organism. Different combinations of canonical motif parameters derived from the analysis of known instances of yeast transcription factor binding sites (length ranging from 6 to l2, number of substitutions from 1 to 4) are automatically tried by the algorithm in different runs. It also analyzes and compares the top-scoring	13

	their occurrences (with substitutions) from input sequences.	motifs of each run with a simple clustering method to detect which ones could be more likely to correspond to transcription factor binding sites. Best instances of each motif are selected from sequences using a weight matrix built with sites found by consensus-based algorithms. http://159.149.109.9/modtools/	
YMF	Uses an exhaustive search algorithm to find motifs with the greatest *z*-scores.	A *p*-value for the *z*-score is used to assess the significance of the motif. Motifs themselves are short sequences over the IUPAC alphabet with spacers ("N"s) constrained to occur in the middle of the sequence. http://bio.cs.washington.edu/software.html#yms	14
Composite Module Analyst (CMA)	Uses a multicomponent fitness function for selection of the promoter model which fits best to the observed gene expression profile.	Defines a promoter model based on composition of transcription factor binding sites and their pairs. Adjusts the results of the fitness function using a genetic algorithm for the analysis of functionally related or coexpressed genes. http://www.gene-regulation.com/cgi-bin/CMA/cma.cgi	15
REDUCE	Motif-based regression method for microarray analysis.	The only required inputs are (i) a single genome-wide set of absolute or relative mRNA abundances and (ii) the DNA sequence of the regulatory region associated with each gene that is probed. REDUCE uses unbiased statistics to identify oligonucleotide motifs whose occurrence in the regulatory region of a gene correlates with the level of mRNA expression. Regression analysis is used to infer the activity of the transcriptional module associated with each motif. http://bussemaker.bio.columbia.edu/reduce/	16
MotifRegressor	Combines the advantages of matrix-based motif finding and oligomer motif-expression regression analysis.	MotifRegressor first constructs candidate motifs and then applies regression analysis to select motifs that are strongly correlated with changes in gene expression. It is particularly effective in discovering expression-mediating motifs of medium-to-long width with multiple degenerate positions. MotifRegressor relies in part on MDScan, a software package developed by the Brutlag Lab at Stanford University. http://www.math.umass.edu/~conlon/mr.html	17
CisModule	Employs a hierarchical mixture approach to model the cis-regulatory module structure.	It is based on the hierarchical mixture model, followed by a *de novo* motif-module discovery algorithm using the Bayesian inference of module locations and within-module motif sites. Dynamic programming-like recursions are developed to reduce the computational complexity from exponential to linear in sequence length. http://www.stat.ucla.edu/~zhou/CisModule/index.html	18

References

1. Hughes JD, Estep PW, Tavazoie S, Church GM (2000). Computational identification of *cis*-regulatory elements associated with functionally coherent groups of genes in *Saccharomyces cerevisiae*. J Mol Biol 296:1205–1214.
2. Workman CT and Stormo GD (2000). ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. PacificSymposium on Biocomputing (ed. Altman R, Dunker AK, Hunter L, Klein TE). 467–478 (Stanford University, Stanford, CA).

3. Hertz GZ and Stormo GD (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15:563–577.
4. Frith MC, Hansen U, Spouge JL, Weng Z (2004). Finding functional sequence elements by multiple local alignment. Nucleic Acids Res 32:189–200.
5. Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE (2004). Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. Science 305:1743–1746.
6. Bailey TL and Elkan C (1995). The value of prior knowledge in discovering motifs with MEME. Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology. 21–29 (AAAI Press, Menlo Park, CA).
7. Eskin E and Pevzner P (2001). Finding composite regulatory patterns in DNA sequences. Bioinformatics (Supplement 1) 18:S354–S363.
8. Thijs G, *et al* (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17:1113–1122.
9. van Helden J, Andre B, Collado-Vides J (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281:827–842.
10. van Helden J, Rios AF, Collado-Vides J (2000). Discovering regulatory elements in noncoding sequences by analysis of spaced dyads. Nucleic Acids Res 28:1808–1818.
11. Régnier M and Denise A (2004). Rare events and conditional events on random strings. Discrete Math Theor Comput Sci 6:191–214.
12. Favorov AV, Gelfand MS, Gerasimova AV, Mironov AA, Makeev VJ (2004). Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length and its validation on the ArcA binding sites. Proceedings of BGRS 2004 (BGRS, Novosibirsk).
13. Pavesi G, Mereghetti P, Mauri G, and Pesole G (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res 32:W199–W203.
14. Sinha S and Tompa M (2003). YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res 31:3586–3588.
15. Konovalova T, Valeev T, Cheremushkin E, Kel AE (2005). Composite Module Analyst: Tool for Prediction of DNA Transcription Regulation. Testing on Simulated Data. ICNC 2:1202-1205.
16. Roven C and Bussemaker HJ (2003). REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. Nucleic Acids Research 31(13):3487-3490.
17. Conlon EM, Liu XS, Lieb JD, Liu JS(2003). Proc Natl Acad Sci USA 100 (6):3339.
18. Zhou Q and Wong WH (2004). CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci USA 101:12114-12119.

Index

B

C

D

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V

W

Y