Illumina® Systems & Software

BeadStudio User Guide

Data Analysis Software for Use with Illumina Gene Expression Products

ILLUMINA PROPRIETARY
Doc. # 11179632 Rev. B

Notice

This publication and its contents are proprietary to Illumina, Inc., and are intended solely for the contractual use of its customers for no other purpose than to operate the system described herein. This publication and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina, Inc.

For the proper operation of this system and/or all parts thereof, the instructions in this guide must be strictly and explicitly followed by experienced personnel. All of the contents of this guide must be fully read and understood prior to operating the system or any of the parts thereof.

FAILURE TO COMPLETELY READ AND FULLY UNDERSTAND AND FOLLOW ALL OF THE CONTENTS OF THIS GUIDE PRIOR TO OPERATING THIS SYSTEM, OR PARTS THEREOF, MAY RESULT IN DAMAGE TO THE EQUIPMENT, OR PARTS THEREOF, AND INJURY TO ANY PERSONS OPERATING THE SAME.

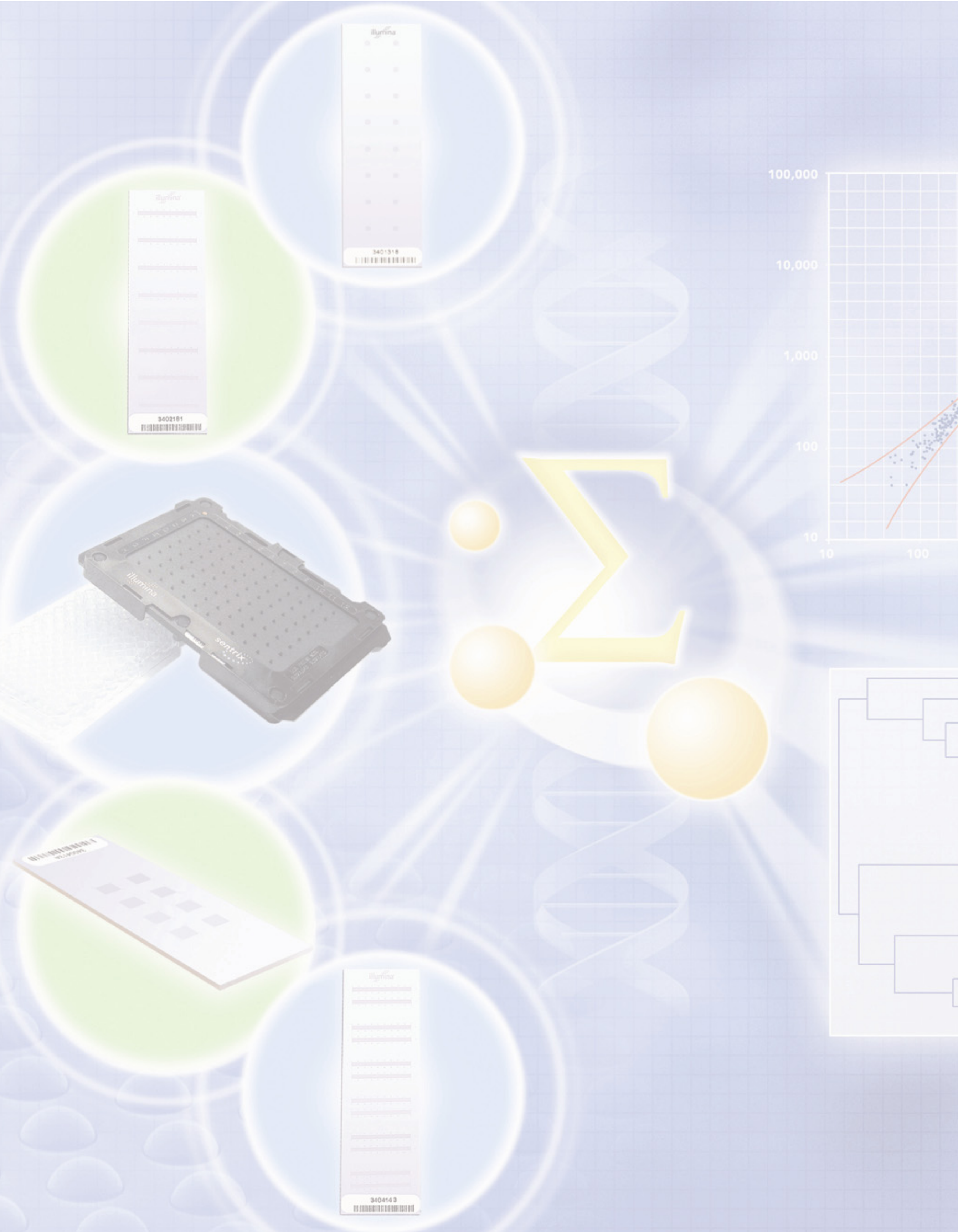
Illumina, Inc. does not assume any liability arising out of the application or use of any products, component parts or software described herein. Illumina, Inc. further does not convey any license under its patent, trademark, copyright, or common-law rights nor the similar rights of others. Illumina, Inc. further reserves the right to make any changes in any products, or parts thereof, described herein without notice.

DASL, GoldenGate, Illumina, Infinium, and Sentrix are trademarks of Illumina, Inc. Other brand and product names mentioned herein may be trademarks or registered trademarks of their respective owners.

© 2004, 2005 Illumina, Inc. All rights reserved.

Revision History

Revision Letter	Release Date
A	12-1-04
B	3-18-05


100,000
10,000
1,000
100
10
10
100
illumina
sentrix

Contents

Chapter 1 Overview

Introduction 1-2
BeadStudio Terminology 1-4
Install BeadStudio 1-5
Start BeadStudio 1-8
Content Descriptor File 1-9

Chapter 2 Experiment Creation & Analysis

Introduction 2-2
Processing Experiments 2-3
Process an Experiment Using the New Experiment Wizard 2-4
Populating Groups 2-8
Removing Groups 2-11
Analyzing Gene Expression Data 2-12
Viewing Your Data 2-17
Process an Experiment Outside the Wizard: Using the BeadStudio Main Page 2-18
Analyzing Gene Expression Data 2-20
Re-Launching the New Experiment Wizard 2-22
Saving Current Experiment & Loading Saved Experiments 2-23
Viewing Your Data 2-24
Shortcut Tools for Defining Experiments 2-25
Set Up & Apply Group Layout Files 2-25
Set Up & Apply Sample Sheet 2-28
Gene Analysis Output Files 2-32
Creating the Mask File 2-40
Browse History 2-41
Clear or Copy Message Log Feature 2-41

Chapter 3 Data Visualization

Introduction 3-2

Accessing the Data Visualization Tools 3-3

From the New Experiment Wizard 3-3

From the BeadStudio Main Page 3-4

Scatter Plots 3-7

Scatter Plot Context Menu Functions 3-16

Finding Genes in the Scatter Plot 3-19

Data Tab 3-23

Manifest Tab 3-25

Ontology Tab 3-27

Other Scatter Plot Functionalities 3-27

Cluster Analysis 3-28

Similarities and Distances 3-29

Analyze Clusters 3-30

Dendrogram Context Menu Selections 3-34

View the Sub-Tree List Directly in the Dendrogram 3-35

Copy/Paste Clusters 3-36

From Scatter Plot to Dendrogram 3-36

From Dendrogram to Scatter Plot 3-38

Control Summary Reports 3-40

Control Summary Reports for DirectHyb 3-40

. . . . 3-42

Control Summary Reports for the DASL™ Assay 3-43

Viewing Control Reports 3-45

Chapter 4 Image Viewing

Introduction 4-2

Using the Image Viewer 4-3

Overlay Cores 4-7

Image Appearance Control 4-8

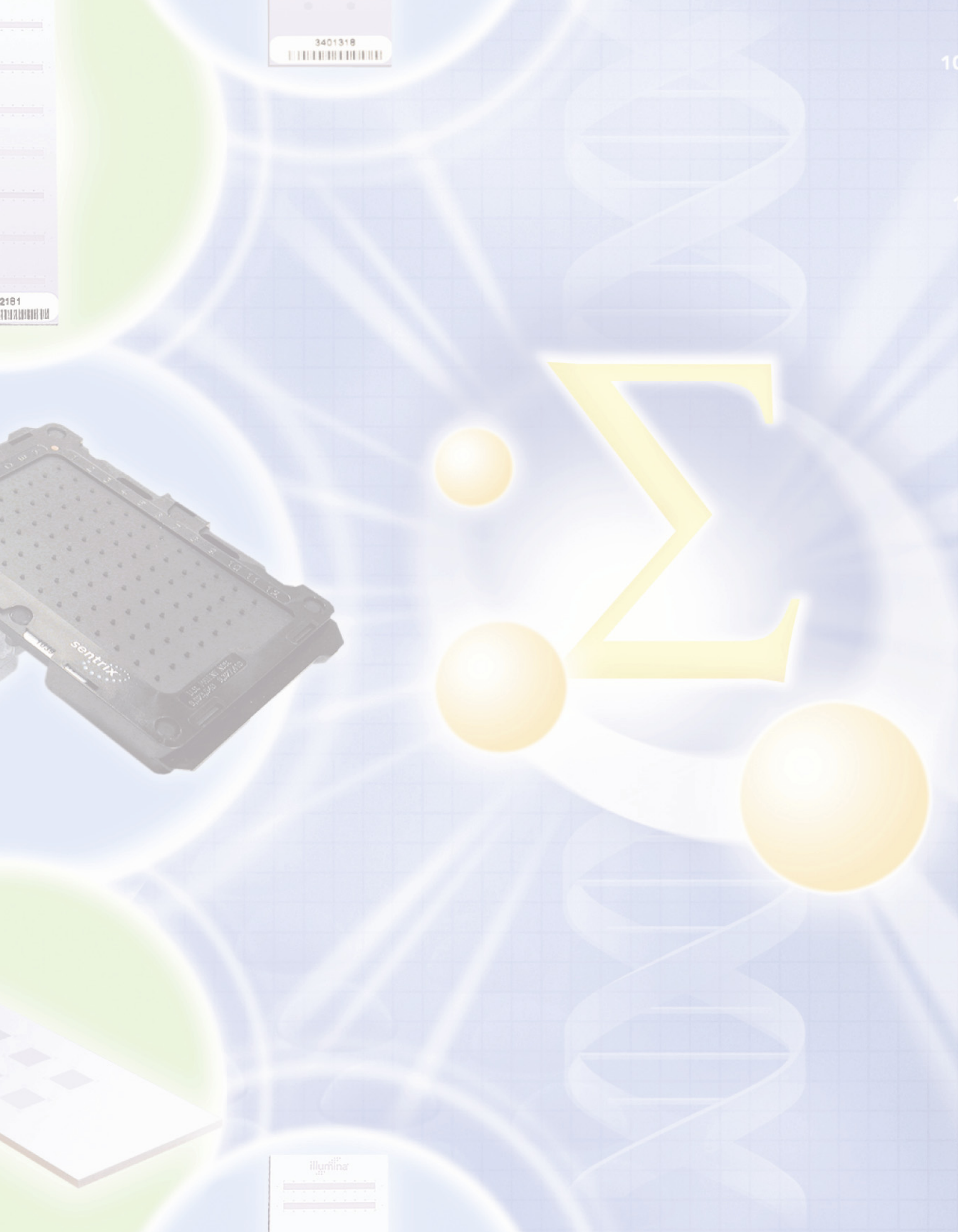
Chapter 5 Screen Elements

BeadStudio User Interface Screen Elements 5-2

BeadStudio Example Screens 5-7

Chapter 6 Normalization & Differential Analysis Algorithms

Introduction 6-2
A Non-Mathematician's Guide 6-2
Normalization Methods 6-2
Definitions 6-3
Background Method 6-3
Average Method 6-4
Cubic Spline Method 6-5
Hyb Controls Method 6-6
Rank-Invariant Method 6-7
Normalization Algorithms 6-8
Background 6-8
Average 6-8
Cubic Spline 6-9
Hyb Controls 6-9
Rank Invariant 6-10
Differential Expression Algorithm 6-11
Illumina Custom 6-11
Mann-Whitney 6-13
T-test 6-14
Detection Score 6-14
Whole Genome BeadChips 6-14
Focused Array & DASL Products 6-15


3401318
2181
sentrix
illumina

Figures

Figure 1-1 Gene Expression Analysis Process Flow 1-3
Figure 1-2 BeadStudio Wizard Welcome Screen 1-5
Figure 1-3 Select Installation Folder Screen 1-6
Figure 1-4 Confirm Installation Screen 1-6
Figure 1-5 Installing BeadStudio Screen 1-7
Figure 1-6 Installation Complete Screen 1-7
Figure 1-7 BeadStudio Application Icon on Your PC Desktop 1-7
Figure 1-8 Welcome Screen 1-8

Figure 2-1 Welcome - Start New Experiment Wizard 2-4
Figure 2-2 BeadStudio Experiment Wizard Welcome Page 2-5
Figure 2-3 Location Page 2-6
Figure 2-4 Samples Page 2-7
Figure 2-5 Groups Pane 2-9
Figure 2-6 Drag and Drop Samples into a Group 2-10
Figure 2-7 Single-Sample per Group 2-11
Figure 2-8 Analysis Page 2-12
Figure 2-9 DASL Mode Analysis Page 2-14
Figure 2-10 Output Parameters Page 2-15
Figure 2-11 How do you want to view your data? Dialog Box 2-17
Figure 2-12 Welcome - Start in the BeadStudio Main Page 2-18
Figure 2-13 BeadStudio Main Page 2-19
Figure 2-14 Main Page Experiment Parameters Dialog Box 2-20
Figure 2-15 Main Page Output Data Selection Dialog Box 2-21
Figure 2-16 Analysis Mode Selected from the Menu Bar 2-22
Figure 2-17 Re-Launch New Experiment Wizard 2-22
Figure 2-18 Save or Save As 2-23
Figure 2-19 Open Experiment 2-23
Figure 2-20 Visualize Data from the Menu Bar 2-24
Figure 2-21 Visualize Data from the Toolbar 2-24
Figure 2-22 Group Layout File Example 2-25
Figure 2-23 Open Pop-Up 2-26
Figure 2-24 Groups Displayed per Applied Group Layout File 2-27

Figure 2-25 *Sample Sheet Examples* 2-30
Figure 2-26 *Open Sample Sheet Pop-Up* 2-31
Figure 2-27 *Missing Sample Sheet Data Files Warning Message* 2-31
Figure 2-28 *Example XXXXXX_gene_profile File* 2-33
Figure 2-29 *Example XXXXXX_gene_probe_profile File* 2-35
Figure 2-30 *Example XXXXXX_gene_diff File* 2-36
Figure 2-31 *Example XXXXXX_qcinfo File for Group Gene Analysis* 2-38
Figure 2-32 *Browse History Pulldown List* 2-41
Figure 2-33 *Message Pane Clear/Copy Button* 2-41

Figure 3-1 *How do you want to view your data? Dialog Box* 3-3
Figure 3-2 *Tools Menu* 3-4
Figure 3-3 *Browse to the Data File* 3-4
Figure 3-4 *Start scatter plot and Start cluster analysis Tool Bar Buttons* . . . 3-5
Figure 3-5 *Scatter Plot Source Data Dialog Box* 3-8
Figure 3-6 *Source Data Dialog Box for Differential Analysis* 3-9
Figure 3-7 *Scatter Plot* 3-10
Figure 3-8 *Scatter Plot Tools Menu* 3-13
Figure 3-9 *Scatter Plot Context Menu* 3-16
Figure 3-10 *Find Genes... Tool Selected* 3-19
Figure 3-11 *Find Genes Dialog Box* 3-20
Figure 3-12 *Zoom in to See Selected Genes* 3-22
Figure 3-13 *Gene Properties: Window Data Tab* 3-23
Figure 3-14 *Gene Properties: Window Data Tab* 3-24
Figure 3-15 *Gene Properties: Window Manifest Tab* 3-25
Figure 3-16 *NCBI Website* 3-26
Figure 3-17 *NCBI Record* 3-26
Figure 3-18 *Gene Properties: Ontology Tab* 3-27
Figure 3-19 *Dendrogram, Similarity Example* 3-29
Figure 3-20 *Dendrogram, Showing Nodes* 3-30
Figure 3-21 *Cluster Analysis Dialog Box* 3-31
Figure 3-22 *Dendrogram* 3-32
Figure 3-23 *Dendrogram with Context Menu* 3-33
Figure 3-24 *Zoom in to View Sub-Tree List* 3-35
Figure 3-25 *Select Region* 3-36
Figure 3-26 *Copy & Paste* 3-37
Figure 3-27 *Select Sub-Tree* 3-38

Figure 3-28 Copy & Paste 3-39

Figure 3-29 Control Summary Report 3-40

Figure 3-30 Housekeeping Controls Secondary Graph 3-41

Figure 3-31 Control Summary Context Menu 3-42

Figure 3-32 Control Summary Reports 3-43

Figure 3-33 Contamination Controls Secondary Graph 3-44

Figure 3-34 Control Summary Context Menu 3-45

Figure 4-1 Right-Click any Sample to Activate Image Viewer 4-3

Figure 4-2 Sample Image 4-4

Figure 4-3 Overlay Cores Image 4-7

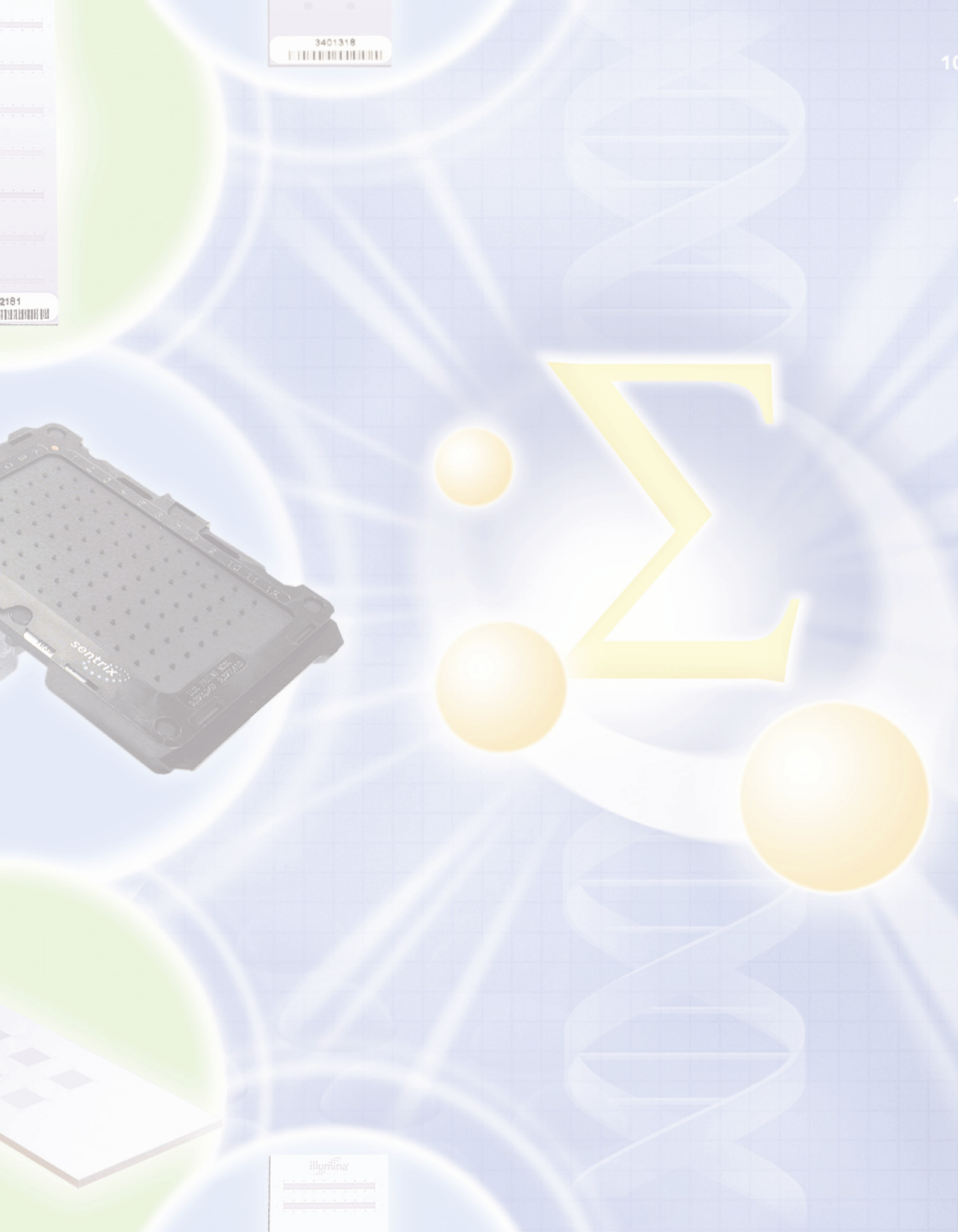
Figure 4-4 Image Control Pane Detail 4-8

Figure 5-1 Extracted Intensity Samples Shown in Tan 5-7

Figure 5-2 Unimaged Samples Shown as Speckled 5-8

Figure 5-3 Extracted Intensities, Problem Samples in Gray 5-8

Figure 6-1 Sample A & Sample B 6-6



Tables

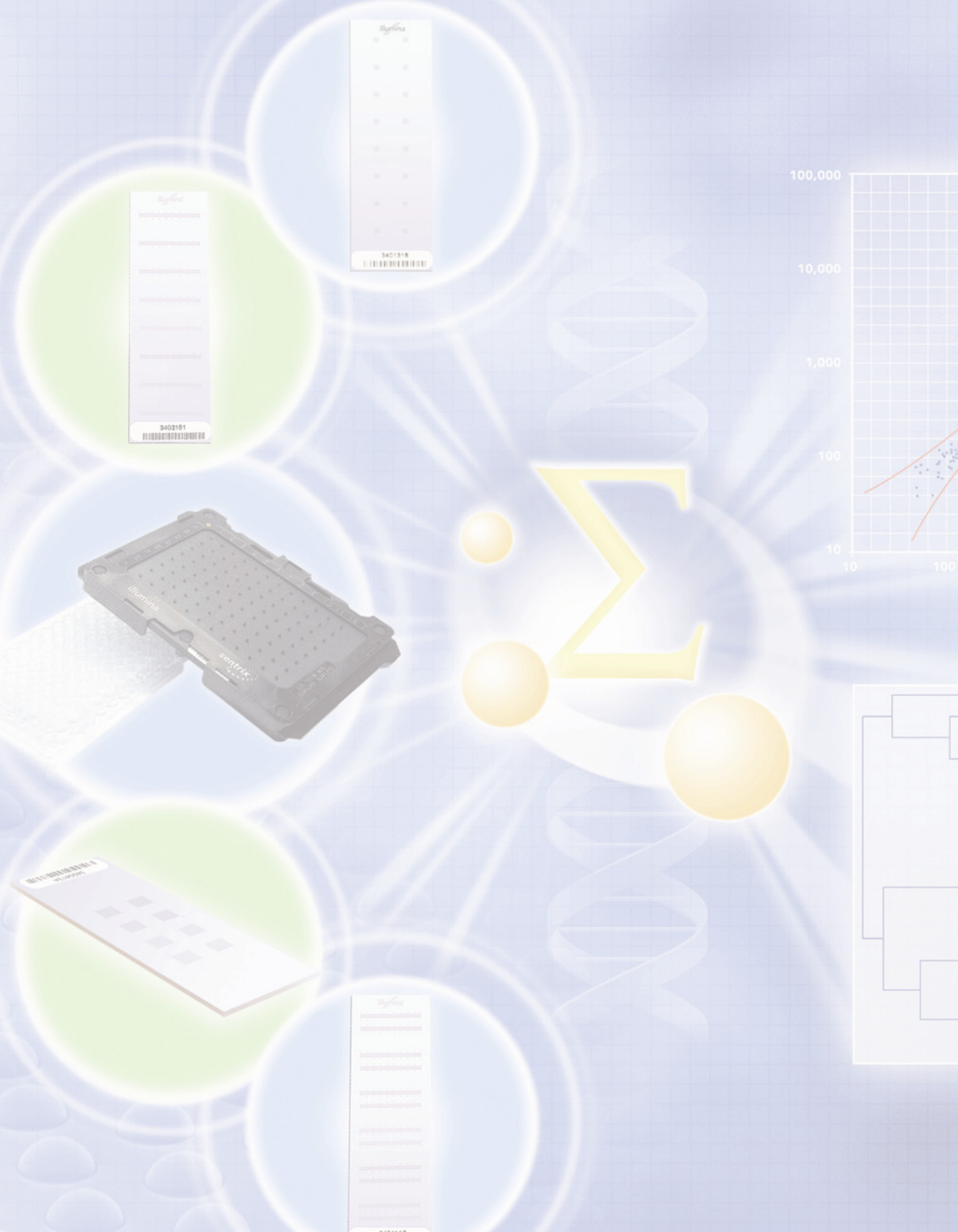
Table 1-1 *BeadStudio Terminology* 1-4
Table 1-2 *Wizard Option Descriptions* 1-9
Table 2-1 *Sample Sheet Guidelines* 2-28

Table 2-2 *Output File Descriptions* 2-32
Table 2-3 *gene_profile Column Descriptions* 2-34
Table 2-4 *gene_diff Column Descriptions* 2-37
Table 2-5 *qcinfo Column Descriptions* 2-39

Table 3-1 *Scatter Plot Control Panel Descriptions* 3-11
Table 3-2 *Scatter Plot Tools Menu Item Descriptions* 3-14
Table 3-3 *Scatter Plot Context Menu Item Descriptions* 3-17
Table 3-4 *Dendrogram Context Menu Descriptions* 3-34

Table 4-1 *Image Viewer Features* 4-5

Table 5-1 *BeadStudio User Interface Screen Elements* 5-2
Table 5-2 *Color Key for Displayed Samples* 5-7


illumina
3401318
3402181
sentrix
100,000
10,000
1,000
100
10
10
100

Chapter 1

Overview

Topics

- Introduction *1-2*
- BeadStudio Terminology *1-4*
- Install BeadStudio *1-5*
- Start BeadStudio *1-8*

Introduction

The BeadStudio software package included with your Illumina® Gene Expression System is a tool for analyzing gene expression data from scanned microarray images collected from the Illumina BeadArray Reader. Resulting BeadStudio files can be used by most standard gene expression analysis programs.

BeadStudio executes two types of data analysis:

- Gene Analysis

  Quantifying gene expression signal levels

- Differential Analysis

  Determining if gene expression levels have changed between two experimental groups

You can perform these analyses on individual samples or on groups of samples treated as replicates.

BeadStudio reports experiment performance based on built-in controls that accompany each experiment.

In addition, BeadStudio provides scatter-plotting and dendrogram tools, facilitating quick, visual means for exploratory analysis.

This chapter describes the BeadStudio software, presents guidelines for evaluating experiment quality, and defines BeadStudio-related terminology.

Wherever this symbol appears, a special note for the Illumina® DASL™ Assay appears in the text.

Figure 1-1 illustrates the gene expression analysis process flow.

Figure 1-1 Gene Expression Analysis Process Flow

BeadStudio Terminology

Terms used in this chapter are listed and defined below.

Table 1-1 BeadStudio Terminology

Term	Definition
Content Descriptor	The collection of probe sequences represented on a SAM or BeadChip. DASL™ Assay: For the DASL Assay, content descriptor refers to the collection of probe sequences in the oligo pool.
Background	An estimate of signal produced by all sources except specific binding of target to probe. This estimate is established by averaging the signal of a large number of negative control probes.
Bead Type	A collection of beads carrying identical probe sequences DASL™ Assay: The DASL Assay uses universal SAMs and BeadChips. Bead Type refers to a collection of beads carrying identical address sequences.
Control category	A group of probes and/or targets with a common functional role (see description of particular control categories)
Error model	A statistical model used to determine p-values for differential expression
Experiment	A collection of groups of samples representing a biological study
Experiment file	An XML file listing group membership and experimental parameters
Group	Samples combined according to some experimental criteria
Normalization	The mathematical process of adjusting intensity values in order to improve the quality of data comparison among samples
Probe	A 50mer oligonucleotide designed to be complementary to a particular target sequence. Probes are immobilized on the surface of glass beads randomly assembled into the SAM or BeadChip. Hybridization of probes provides intensity values for analysis of target abundance. DASL™ Assay: For the DASL Assay, a probe refers to the oligo sequences that hybridize to cDNA target sequences. These oligos also contain address sequences that hybridize to sequences immobilized on the surface of glass beads randomly assembled into a universal SAM or BeadChip. Hybridization of DASL Assay products provides intensity values for analysis of target abundance.
Reference group	A group to which other experimental groups are compared during differential analysis
Repository	A folder on your computer or network where scanned SAM or BeadChip images and raw intensity files are stored

Install BeadStudio

Follow the steps below to properly install the BeadStudio application.

1. Double-click on “setup.exe” in the BeadStudio directory of the Installation CD you received with your system and follow the onscreen instructions.

2. The installation process will install the Microsoft .NET Framework on your system if it has not already been installed. Installation of the .NET Framework requires administrative privileges and may take several minutes.

The BeadStudio Setup Wizard will guide you through the installation process (Figures 1-2 thru 1-7).

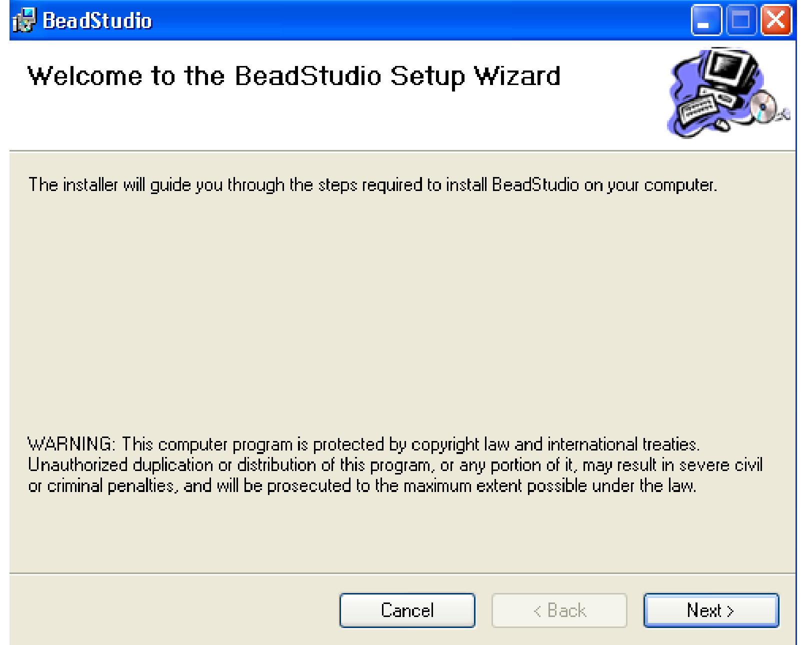


Figure 1-2 BeadStudio Wizard Welcome Screen

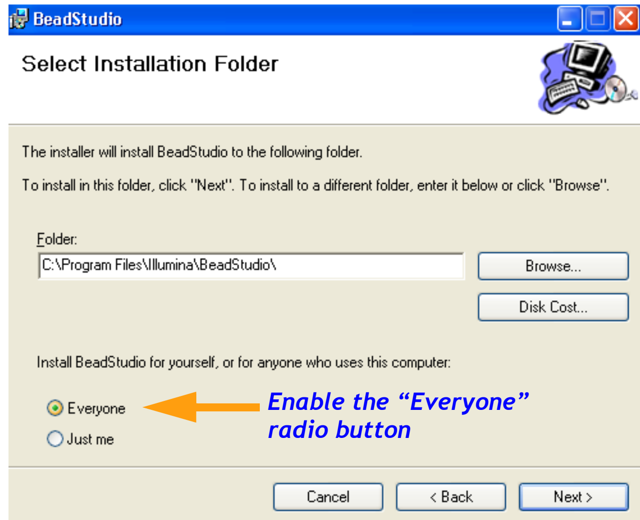


Figure 1-3 Select Installation Folder Screen

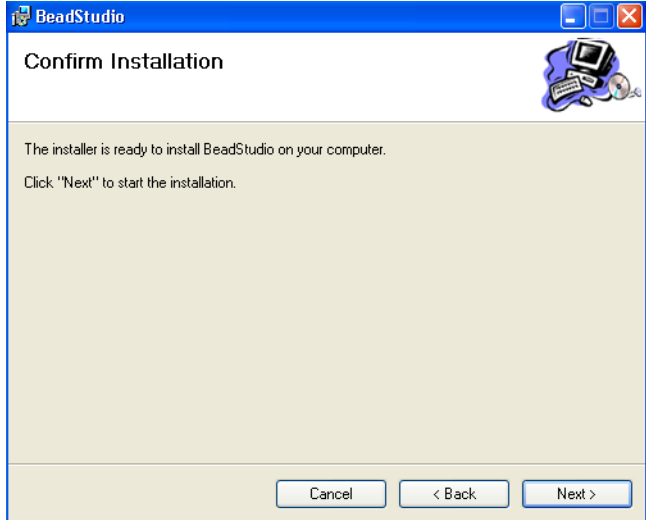


Figure 1-4 Confirm Installation Screen

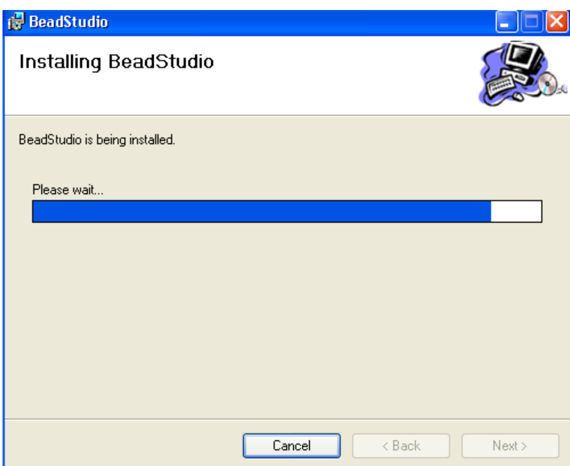


Figure 1-5 Installing BeadStudio Screen

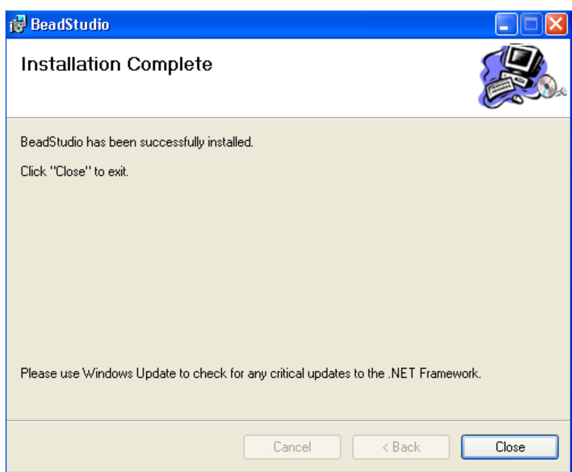


Figure 1-6 Installation Complete Screen

3. The BeadStudio application icon (Figure 1-7) now appears on your PC desktop.

Figure 1-7 BeadStudio Application Icon on Your PC Desktop

Start BeadStudio

1. On your PC desktop, double-click the BeadStudio icon to start the application.
2. Select wizard and mode:
   a. At the **Welcome** screen (Figure 1-8), click the appropriate radio buttons in the **What would you like to do?** pane. Table 1-2 describes the options.

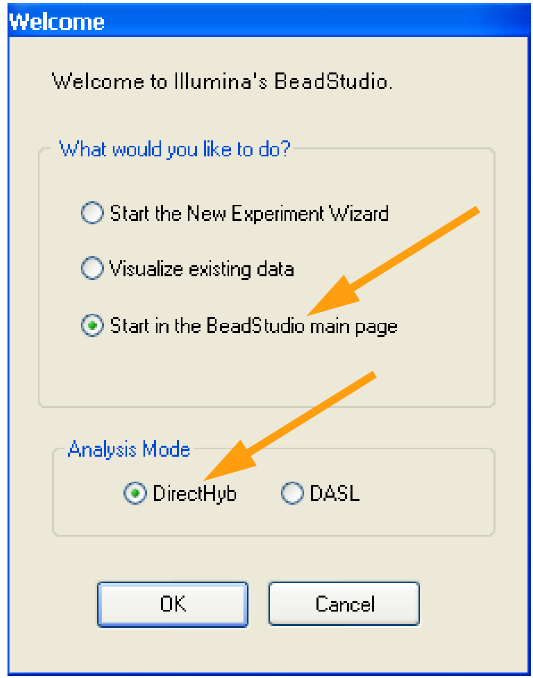


Figure 1-8 Welcome Screen

Table 1-2 Wizard Option Descriptions

Option	Description
Start the New Experiment Wizard	This wizard walks you through the steps to: - Select which samples to analyze - Select a normalization method for your data - Generate output files containing gene expression data. ▷ *Use these output files for further analysis in BeadStudio, such as scatter plotting, clustering (dendrograms), and generating experimental quality reports (control summary reports) (all described in Chapter 3).*
Start the Visualize Existing Data Wizard	Select this option for experiments you have already processed through the BeadStudio application. Use this wizard to explore the data through scatter plotting, cluster analysis (dendrograms), or by generating experimental quality reports (control summary reports) (all described in Chapter 3). If you have not already processed an experiment, run the Start New Experiment Wizard first.
Start in the BeadStudio Main Page	BeadStudio functions may be carried out in a less structured format, without using a wizard. In this mode you can also modify existing experiments, or repeat their analysis using different normalization or output file format settings.

b. If you have purchased both the *DirectHyb* and *DASL™ Gene Expression Systems*, the **Analysis Mode** pane displays a choice. Click the radio button to select your mode (see Figure 1-8). If you have purchased only one of these systems, the **Analysis Mode** pane will be grayed out and the appropriate mode will be automatically selected.

3. Click **OK**.

Content Descriptor File

The BeadStudio application uses a **Content Descriptor** file (.xml) to map BeadTypes to targets. Illumina provides this file on a CD that you receive with each Sentrix Array Matrix or BeadChip.

NOTES

Chapter 2

Experiment Creation & Analysis

Topics

- Introduction *2-2*
- Processing Experiments *2-3*
  - *Process an Experiment Using the New Experiment Wizard 2-4*
  - *Process an Experiment Outside the Wizard: Using the BeadStudio Main Page 2-18*
- Shortcut Tools for Defining Experiments *2-25*
  - *Set Up & Apply Group Layout Files 2-25*
  - *Set Up & Apply Sample Sheet 2-28*
- Creating the Mask File *2-40*
- Browse History *2-41*
- Clear or Copy Message Log Feature *2-41*

Introduction

Using the intensity files produced by the BeadArray Reader, BeadStudio's *Gene Analysis* tool produces output files containing:

- Probe and gene lists
- Associated hybridization intensities (normalized or raw)
- Information about the system controls

If desired, BeadStudio's *Differential Analysis* tool can produce output files determining the probability that a gene's signal has changed between two samples or groups of samples.

Using these output files, BeadStudio's *Data Visualization* tools can create more sophisticated plotting analyses such as Scatter Plots, Cluster Analysis Dendrograms, and Control Summary Graphs.

To produce the BeadStudio output files, you first define an experiment. In a BeadStudio experiment, you define the samples used and their grouping (sample sets that can be compared against each other for the purpose of identifying gene expression differences).

To define your experiment, first specify your groups, then assign samples to them. In the simplest experiment, each group will have only one sample. However, if your experiment includes replicate samples, you can assign these to the same group. Within a group, BeadStudio will average the values for each gene across the samples, and its algorithms will automatically take advantage of the replicates' statistical power to provide more sensitive determination of detection and differential expression.

Processing Experiments

The next two sections, *Process an Experiment Using the New Experiment Wizard* (page 2-4) and *Process an Experiment Outside the Wizard: Using the BeadStudio Main Page* (page 2-18), provide step-by-step instructions on how to:

- Define an experiment
- Create groups
- Select normalization and differential expression options
- Select fields for, then generate output files
- Plot and cluster output file data

The New Experiment Wizard guides you through the process; the BeadStudio Main Page allows you to carry out the same functions in a less structured format. See Table 1-2, *Wizard Option Descriptions* on page 1-9 for more information.

Process an Experiment Using the New Experiment Wizard

1. After starting the BeadStudio application (see *Start BeadStudio* on page 1-8), at the **Welcome** screen **What would you like to do?** pane, select the radio button for Start the New Experiment Wizard (Figure 2-1).
2. At the **Analysis Mode** pane, click the radio button to select the appropriate mode.

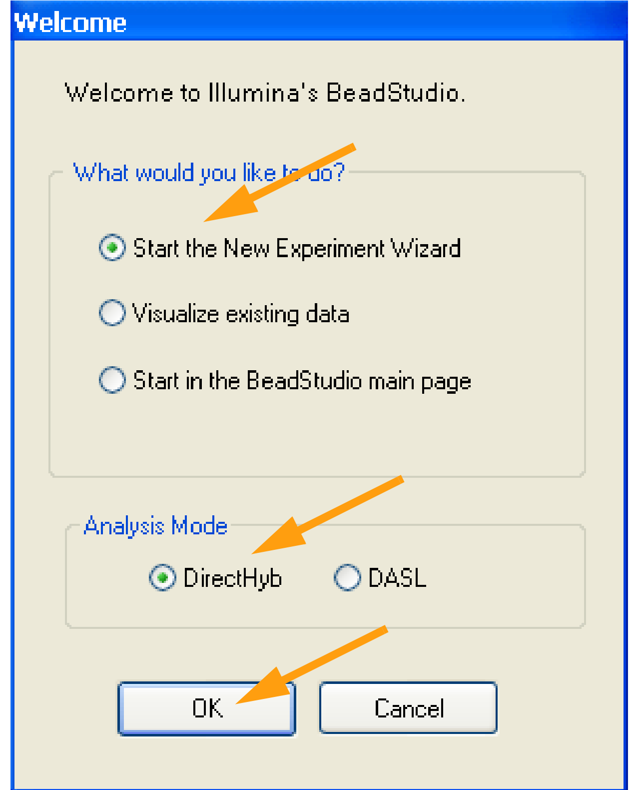


Figure 2-1 Welcome - Start New Experiment Wizard

3. Click **OK.**

4. At the BeadStudio Experiment Wizard Welcome Page (Figure 2-2), click **Next**.

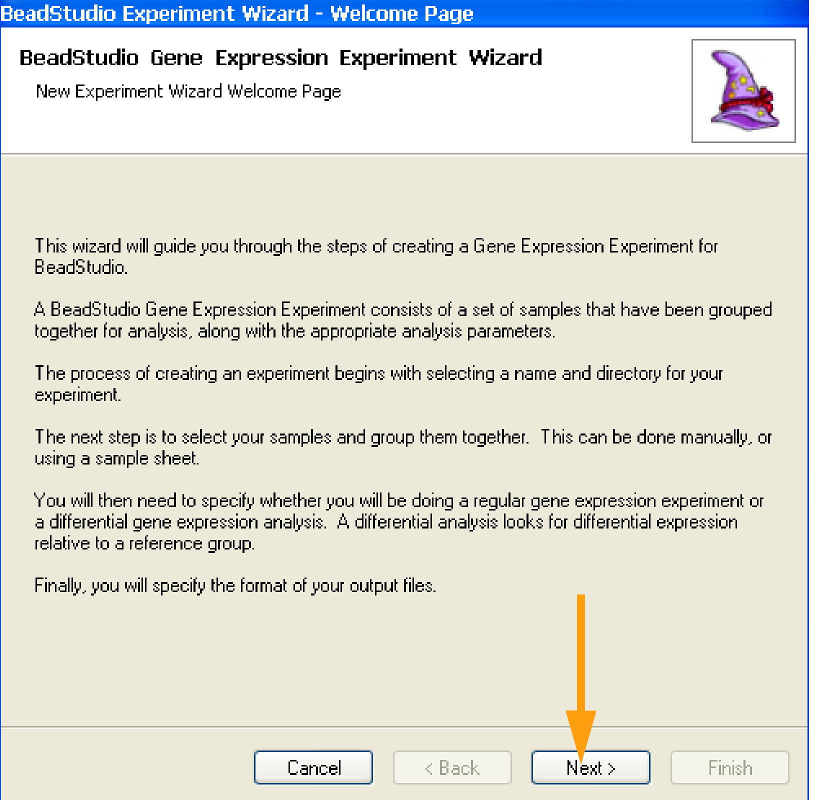


Figure 2-2 BeadStudio Experiment Wizard Welcome Page

5. At the Location Page (Figure 2-3):

   a. Enter a name for your experiment in the **New Experiment Name** field.

   b. Enter or browse for the location to which you want to save your experiment.

NOTE:

Although not required, Illumina recommends storing all your experiments in a common location on your local hard drive or network.

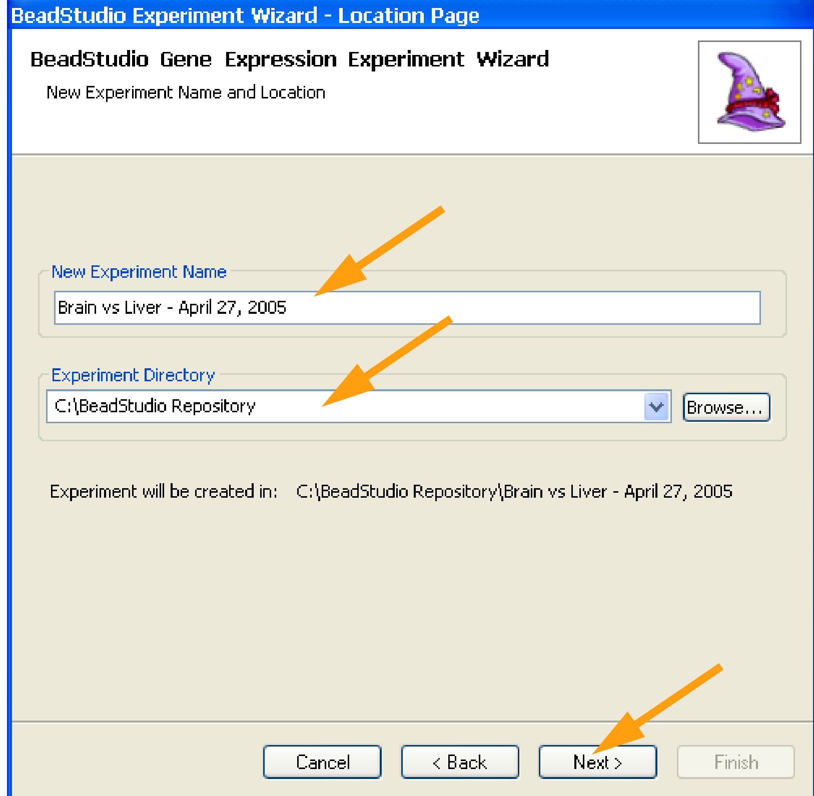


Figure 2-3 Location Page

6. Click **Next**.

7. At the Samples Page (Figure 2-4), in the **How would you like to select your samples and groups?** pane, click the radio button next to your choice.

NOTE:

This section describes the manual method for selecting samples and groups. Optionally, you can also use the Sample Sheet and Group Layout file to help create your experiments in BeadStudio. You can create these files using the example templates provided on your BeadStudio CD, Illumina part # 11182401, provided with your system.

You can also use the Bead Manifest file (optionally) to include more annotated data in your analysis. This file resides on the CD you received with your BeadChips or SAMS.

For the DASL Assay, this file resides on the CD you received with your oligo pool (DAP).

For information on using Sample Sheets, see *Set Up & Apply Sample Sheet* on page 2-28.

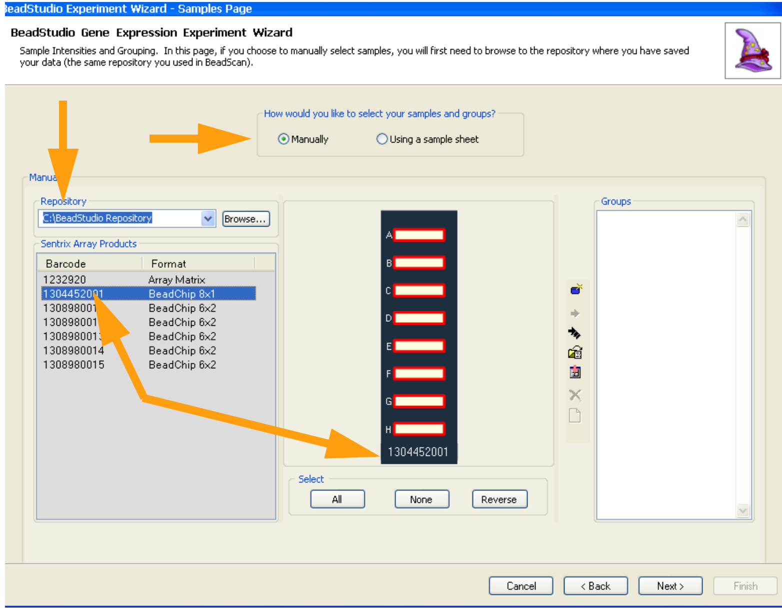


Figure 2-4 Samples Page

8. In the **Repository** pane, enter or browse to the directory in which you have your images and raw data files from the BeadArray Reader.

9. Click on a barcode in the **Select Barcodes** pane to display that BeadChip or SAM in the middle pane.

Populating Groups

This section describes the four methods for populating groups:

- *Standard Method (see below)*
- *Single-Sample per Group (see page 2-11)*
- *Sample Sheet Method (see page 2-11)*
- *Group Layout Method (see page 2-11)*

Standard Method

1. From the displayed BeadChip or SAM, select the samples you want in your groups.
    - *Selected samples are outlined in* ***red***
    - *Initially, all samples are selected*
    - *To de-select samples, click* ***None*** *in the* ***Select*** *pane*
    - *To select an individual sample, click it*
    - *To select multiple samples, Ctrl-click each desired sample*

2. At the **Groups** toolbar, click **New group...** to create a new group.

3. In the **Groups** pane (Figure 2-5), rename “Group 1” to the name of your choice.

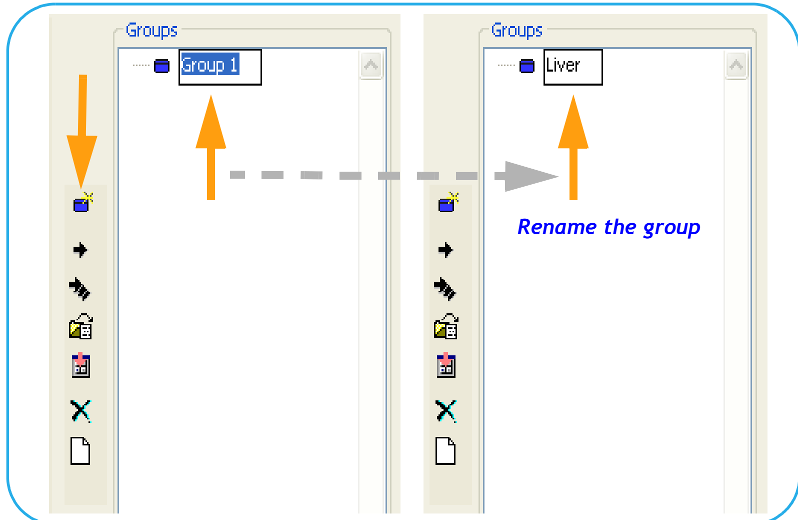


Figure 2-5 Groups Pane

4. To add samples to this group, either:
   - *From the **Select** pane, select them, then click **Add to Group*** .

   OR

   - *Drag and drop the selected samples from the **Select** pane to the name icon of the desired group (Figure 2-6).*

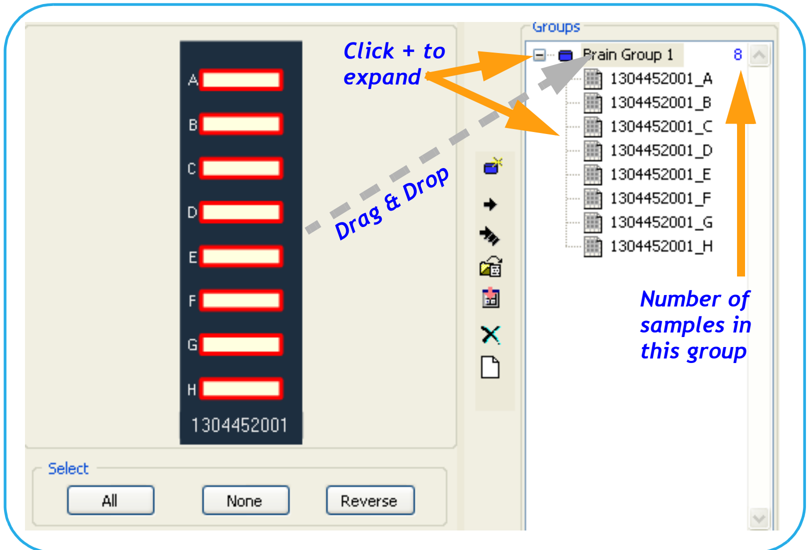


Figure 2-6 Drag and Drop Samples into a Group

5. To populate groups with samples from multiple BeadChips or SAMs, highlight the barcode in the **Repository** pane and select the desired samples.
6. The number to the right of each group icon is the number of samples contained in that group. To expand the group tree and view see the sample identities, click the **+** sign to the left of the group name.

Single-Sample per Group

Highlight the samples you wish to add in the **Select** pane, then click **Create a group for each sample** .

A group will be created for each selected sample (Figure 2-7). Each group is named for the sample it contains. Groups from additional BeadChips or SAMs may be added to the same experiment by the method described in *Populating Groups* on page 2-8, step 5.

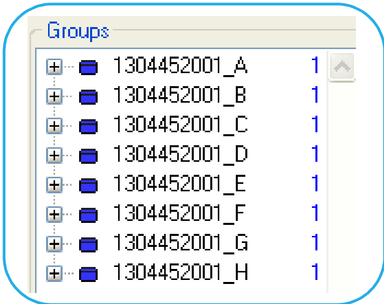


Figure 2-7 Single-Sample per Group

Sample Sheet Method

See *Shortcut Tools for Defining Experiments* on page 2-25.

Group Layout Method

See *Shortcut Tools for Defining Experiments* on page 2-25.

Removing Groups

To remove a selected group, click **Remove...** .

To remove all groups, click **Clear all groups...** .

Analyzing Gene Expression Data

Once you have created your groups, you have fully defined your experiment. At the Samples Page, click Next > to access the **Analysis Page** (Figure 2-8).

1. In the **Analysis** pane, you may select **Gene Expression** or **Differential Expression.**

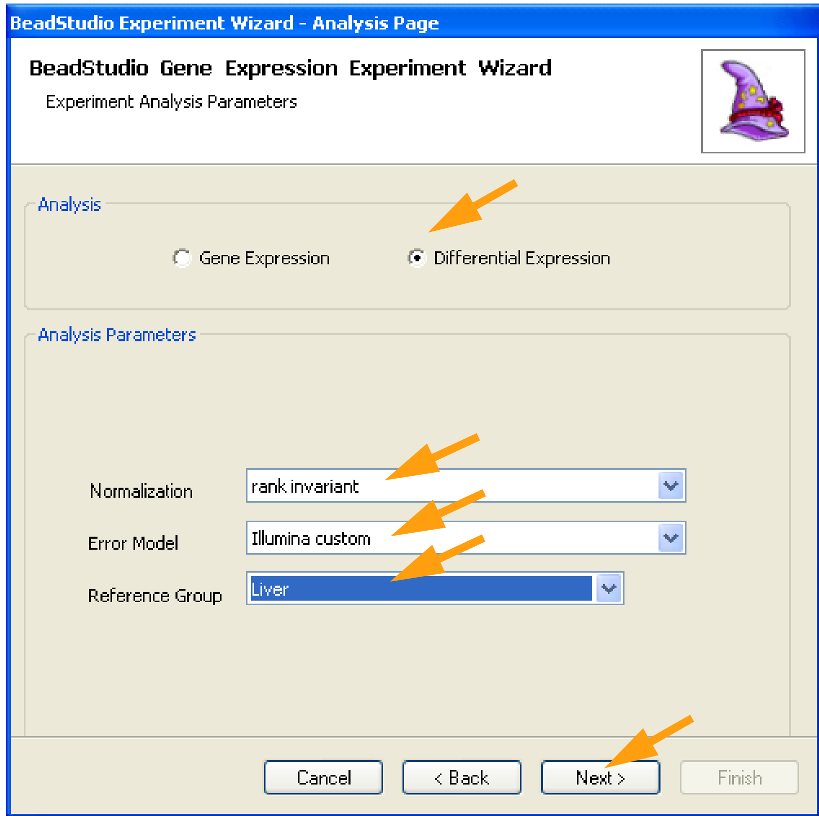


Figure 2-8 Analysis Page

NOTE:

If you select **Gene Expression**, BeadStudio will generate a file containing gene expression intensity values for each gene in each experimental group, along with various other quality-related measurements.

If you select **Differential Expression**, BeadStudio will generate a file with the same information as for Gene Expression, including a differential expression score for each gene in each group. This score represents the statistical confidence that the gene's expression has changed with respect to a reference group.

2. In the **Analysis Parameters** pane (Figure 2-8), use the pulldown menus to select the **Normalization** method, **Error Model**, and **Reference Group**. (See Table 1-1, *BeadStudio Terminology* on page 1-4 for definitions of these terms.)

NOTE:

Normalization method applies to both Gene Expression and Differential Expression. Error Model and Reference Group apply only to Differential Expression.

For most routine whole genome expression experiments, Illumina recommends the **rank invariant** normalization method.

For Differential Expression experiments, we recommend the **Illumina custom** Error Model.

For Differential Expression, choose any of your defined groups as the **Reference Group**.

See Chapter 6 for complete lists and descriptions of Normalization methods and Error Models.

The DASL mode **Analysis Page** has two additional parameter options in the **Analysis Parameters** pane (Figure 2-9):

- Content Descriptor file

  *An .xml file describing the gene content of your DAP oligo pool. BeadStudio uses the **Content Descriptor file** to provide biological information for the sample signals. This file is named according to the convention XXXX-DAP.xml, where XXX is an Illumina-assigned barcode. For further information, see Content Descriptor File on page 1-9.*

- Mask file

  *Allows you to choose which probes to include in your analysis. If you have created a **Mask file**, check the **Use Mask File** box, then browse to the location of this file. For more information on Mask files, please see Creating the Mask File on page 2-40.*

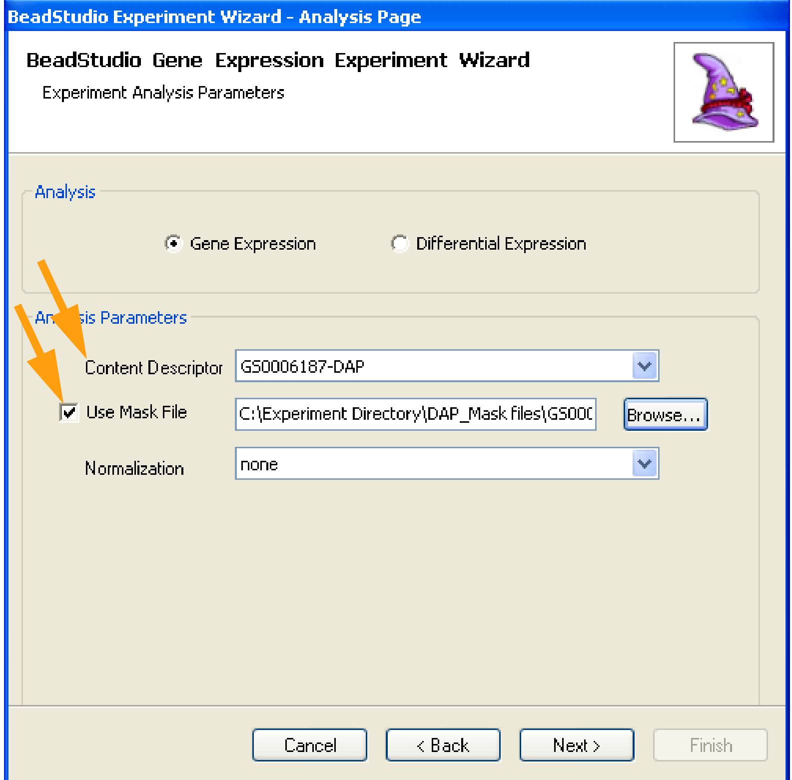


Figure 2-9 DASL Mode Analysis Page

3. Click to access the **Output Parameters Page** (Figure 2-10).

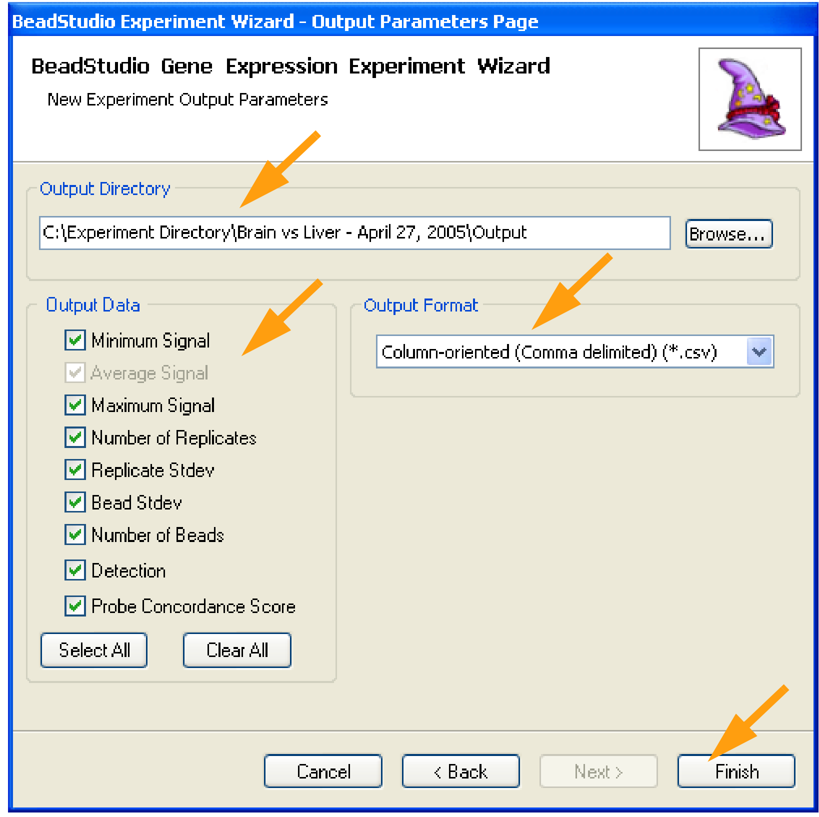


Figure 2-10 Output Parameters Page

4. In the **Output Directory** text field, enter or browse to the location to which you want to save your output data.

NOTE:

The default location is the Experiment Directory you specified at the start of the Wizard.

5. Select the data fields you want the output file to save by checking or un-checking the **Output Data** pane boxes. Tables 2-1 thru 2-5 (pages 2-28 thru 2-39) list and describe these output file data fields.
6. Specify output file format from the **Output Format** pulldown menu.
7. Click .
8. If this is the first time you are using this kind of Matrix (SAM or BeadChip) for DirecyHyb experiments, BeadStudio will prompt for its Content Descriptor (.xml) file. This file associates the BeadArray Reader's scanned image data with the SAM or BeadChip's probe content (see *Content Descriptor File* on page 1-9 for more information). Browse to the file; BeadStudio will load it onto your hard drive.

NOTE:
Once you have loaded the Content Descriptor file onto a given computer, you will never have to load it again.

9. Because different oligo pools could be used for different samples on the same universal SAM or BeadChip, you must browse to the correct Content Descriptor file associated with the oligo pool used for the analyzed samples. If this is the first time you are using an oligo pool (DAP), browse to the file. BeadStudio will load it onto your hard drive. All previously loaded DAP.xml files are available in the **Content Descriptor** pulldown menu (Figure 2-9).
10. As BeadStudio generates and saves your output files to the location you specified, you can observe real-time progress of these operations in the **Message** pane on the **Main Page**.

Viewing Your Data

1. After the output files are generated and saved, select a radio button in the **How do you want to view your data?** dialog box (Figure 2-11). Scatter Plots, Cluster Analysis Dendrograms, and Control Summary Graphs are described in Chapter 3, *Data Visualization*.

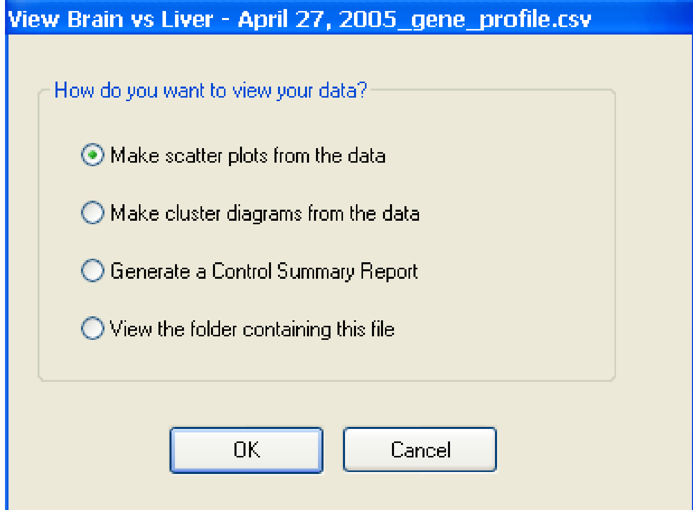


Figure 2-11 How do you want to view your data? Dialog Box

NOTE:

Alternatively, you may wish to open the output files directly. If so, select the **View the folder containing this file** option. to open a Windows folder containing the files. You may view the files in Microsoft Excel or a variety of other applications.

Process an Experiment Outside the Wizard: Using the BeadStudio Main Page

1. You may access the BeadStudio **Main Page** in several ways:

   - *After starting the BeadStudio application (see Start BeadStudio on page 1-8), at the* ***Welcome*** *screen* ***What would you like to do?*** *pane, select the radio button for* ***Start in the BeadStudio Main Page*** *(Figure 2-12).*
   - *After you have run the New Experiment Wizard, at the* ***How do you want to view your data?*** *dialog box (Figure 2-11), click* Cancel *to return to the* ***Main Page****.*
   - *After you have run the New Experiment Wizard and completed data visualization, BeadStudio will automatically display the* ***Main Page****.*

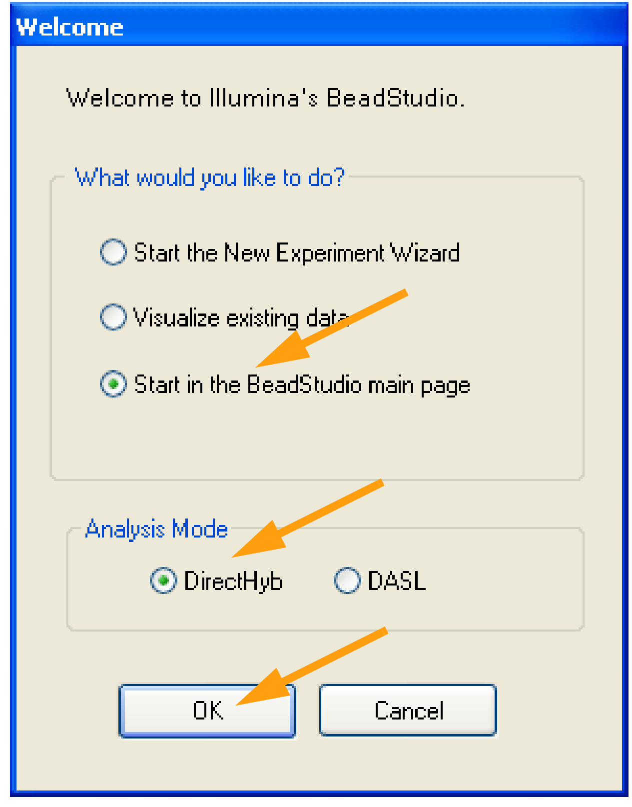


Figure 2-12 Welcome - Start in the BeadStudio Main Page

2. At the **Analysis Mode** pane, click the radio button to select the appropriate mode.
3. Click **OK** to access the **Main Page** (Figure 2-13).

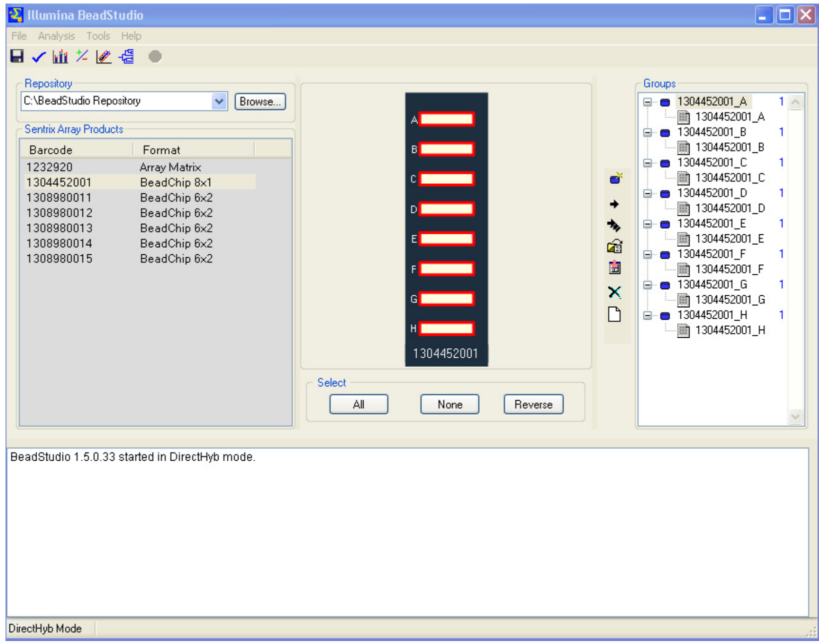


Figure 2-13 BeadStudio Main Page

4. In the **Repository** pane, enter or browse to the directory in which you have your images and raw data files from the BeadArray Reader.
5. Click on a barcode in the **Select Barcodes** pane to display that BeadChip or SAM in the middle pane.
6. Create and populate groups, following the steps in the *Populating Groups* section, starting on page 2-8.

Analyzing Gene Expression Data

Once you have created your groups, you have fully defined your experiment. Use one of two methods to analyze your data using the BeadStudio **Main Page**:

Toolbar Buttons

1. From the main toolbar, click the **Verify Matrix/BeadChip** button . BeadStudio will evaluate the quality of the samples contained on the currently selected SAM or BeadChip. Samples that meet the verification criteria will turn **green**; those that do not meet the criteria will turn gray.

NOTE:

You may review the gray samples and decide either to include or exclude them from BeadStudio's analysis.

2. From the main toolbar, click the **Gene analysis** button or the **Diff analysis** button to launch the **Experiment Parameters** dialog box (Figure 2-14).

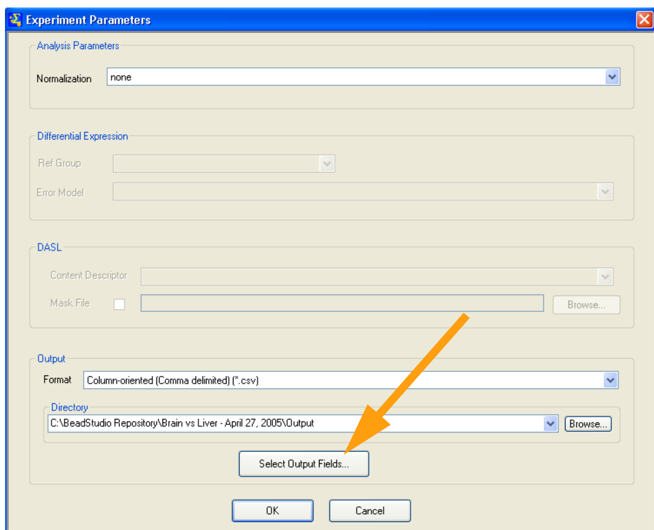


Figure 2-14 Main Page Experiment Parameters Dialog Box

3. Select appropriate parameters, then click **Select Output Fields...**. Select Output Fields... .
4. At the **Output Data Selection** dialog box (Figure 2-15), select the data fields you want the output file to save by checking or un-checking the **Output Data** pane boxes. Table 2-2 on page 2-32 lists and describes these output file data fields.

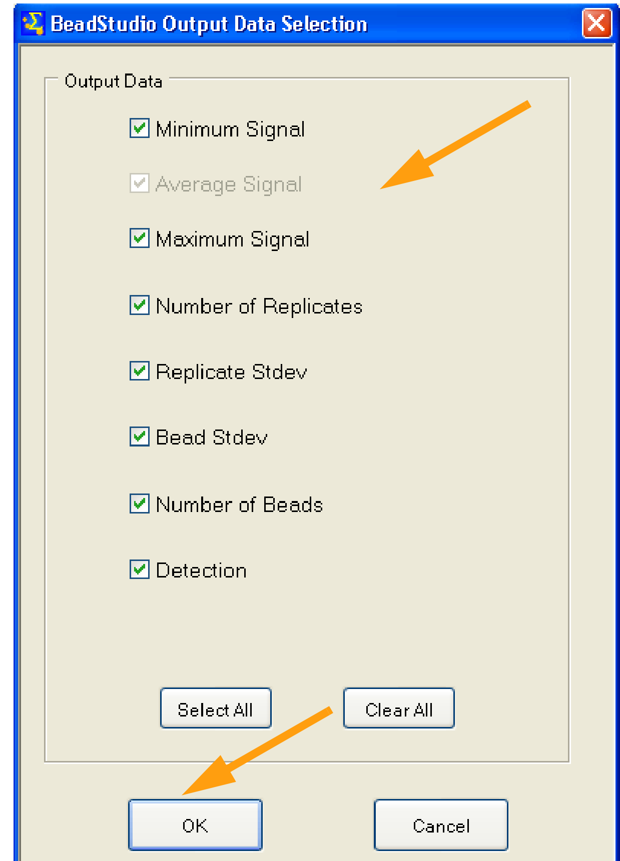


Figure 2-15 Main Page Output Data Selection Dialog Box

5. Click **OK** (BeadStudio will return to the **Experiment Parameters** dialog box).

6. Click **OK** again at the **Experiment Parameters** dialog box to start the output file generation process.
7. From this point, continue from step 8 on page 2-16.

Pulldown Menus

From the main toolbar, select **Analysis | Gene Analysis...** or **Analysis | Differential Expression Analysis** (Figure 2-16).

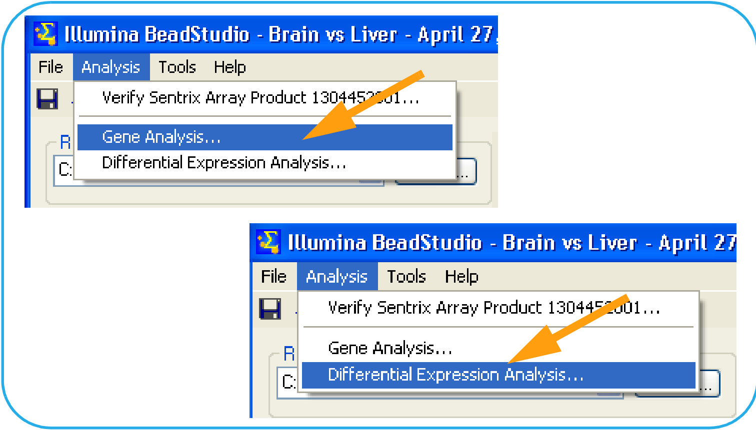


Figure 2-16 Analysis Mode Selected from the Menu Bar

Re-Launching the New Experiment Wizard

If you have been working in the BeadStudio Main Page, and would like to start another new experiment using the New Experiment Wizard (without shutting down and restarting the whole application), from the menu bar, select File | New Experiment Wizard... (Figure 2-17).

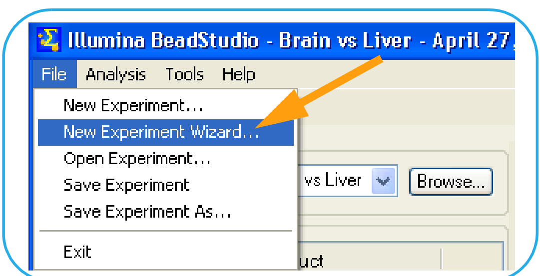


Figure 2-17 Re-Launch New Experiment Wizard

Saving Current Experiment & Loading Saved Experiments

- To save a current experiment, from the menu bar, select **File |Save Experiment** or **File |Save Experiment as….** ().

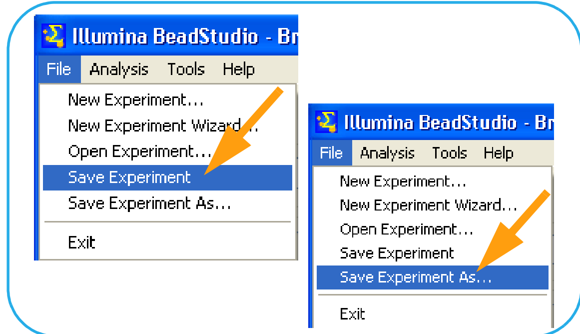


Figure 2-18 Save or Save As

NOTE:

When you save an experiment, BeadStudio creates a file containing the list of selected samples and their groupings. This file is distinct from the gene_profile, gene_diff and qcinfo files generated when you process an experiment through the **Gene Analysis** or **Differential Expression Analysis** tools. By saving an experiment, you can preserve a record of its layout.

- To load a saved experiment and re-process using different options (e.g., different normalization), select **File |Open Experiment…** (Figure 2-19).

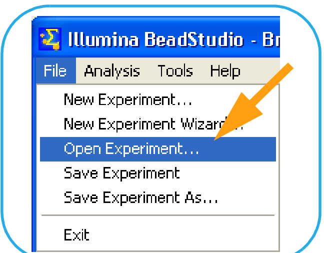


Figure 2-19 Open Experiment

Viewing Your Data

After the output files are generated and saved, to view your data from the **Main Page**:

- From the menu bar, select **Tools |Visualize Existing Data...** (Figure 2-20).

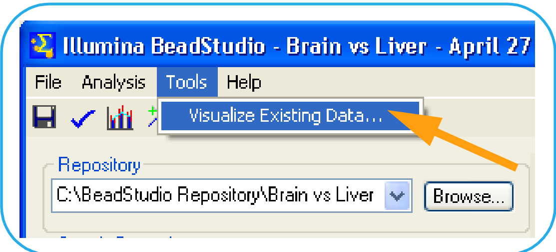


Figure 2-20 Visualize Data from the Menu Bar

- From the toolbar, click **Start scatter plot** or **Start cluster analysis** .

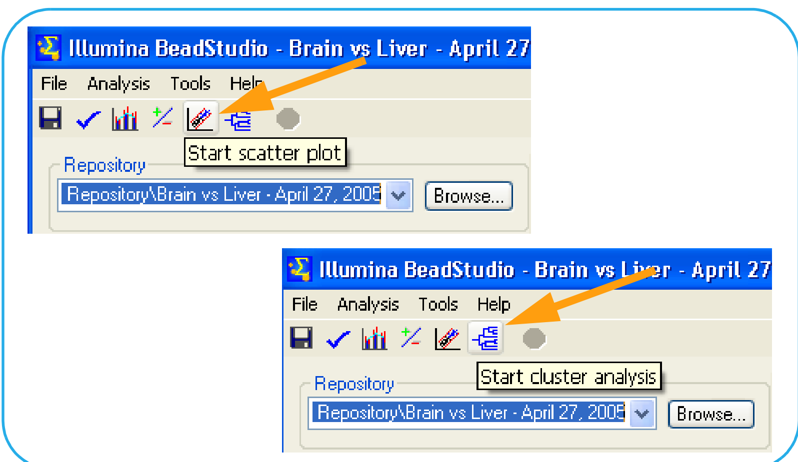


Figure 2-21 Visualize Data from the Toolbar

Scatter Plots, Cluster Analysis Dendrograms, and Control Summary Graphs are described in Chapter 3, *Data Visualization*.

Shortcut Tools for Defining Experiments

Illumina provides two shortcut tools for defining experiments in BeadStudio:

- Group Layout File

  Most useful for the 8 x 12 SAM format, where an experiment is likely to be confined to one SAM. The Group Layout file can be applied to multiple matrices having the same pattern.

- Sample Sheet

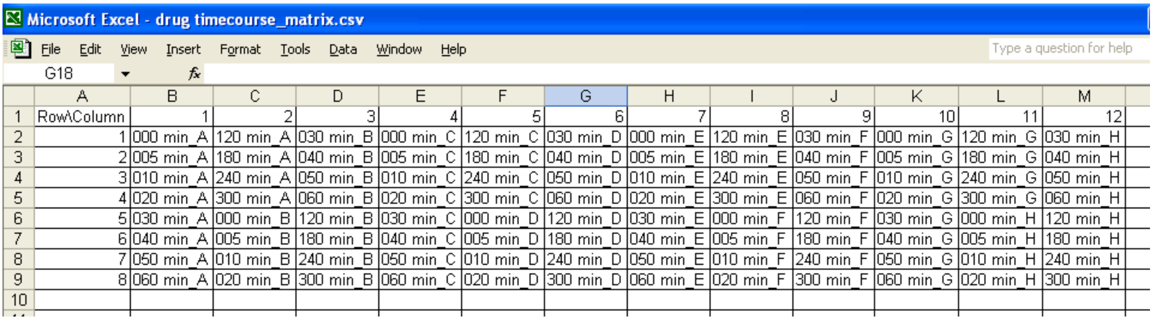
  Most useful for Multiple SAMs or for BeadChips (all formats), where an experiment likely spans several BeadChips.

Set Up & Apply Group Layout Files

BeadStudio provides a quick alternative method for creating large numbers of groups in complex experiments, saving setup time.

The **Apply Group Layout...** option allows you to apply a layout file you have already created (in MS Excel) to one BeadChip or SAM. BeadStudio automatically creates groups according to the applied layout file, then adds selected samples to those groups.

Use the following template as a guide for creating your own group layout files (Figure 2-22). This file must be saved as a *.csv file.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Row\Column	1	2	3	4	5	6	7	8	9	10	11	12
2	1	000 min_A	120 min_A	030 min_B	000 min_C	120 min_C	030 min_D	000 min_E	120 min_E	030 min_F	000 min_G	120 min_G	030 min_H
3	2	005 min_A	180 min_A	040 min_B	005 min_C	180 min_C	040 min_D	005 min_E	180 min_E	040 min_F	005 min_G	180 min_G	040 min_H
4	3	010 min_A	240 min_A	050 min_B	010 min_C	240 min_C	050 min_D	010 min_E	240 min_E	050 min_F	010 min_G	240 min_G	050 min_H
5	4	020 min_A	300 min_A	060 min_B	020 min_C	300 min_C	060 min_D	020 min_E	300 min_E	060 min_F	020 min_G	300 min_G	060 min_H
6	5	030 min_A	000 min_B	120 min_B	030 min_C	000 min_D	120 min_D	030 min_E	000 min_F	120 min_F	030 min_G	000 min_H	120 min_H
7	6	040 min_A	005 min_B	180 min_B	040 min_C	005 min_D	180 min_D	040 min_E	005 min_F	180 min_F	040 min_G	005 min_H	180 min_H
8	7	050 min_A	010 min_B	240 min_B	050 min_C	010 min_D	240 min_D	050 min_E	010 min_F	240 min_F	050 min_G	010 min_H	240 min_H
9	8	060 min_A	020 min_B	300 min_B	060 min_C	020 min_D	300 min_D	060 min_E	020 min_F	300 min_F	060 min_G	020 min_H	300 min_H
10													

Figure 2-22 Group Layout File Example

1. From the **Groups** pane toolbar, click **Apply Group Layout...** .

2. When the **Open** pop-up appears, navigate to the appropriate file and click **Open** (Figure 2-23).

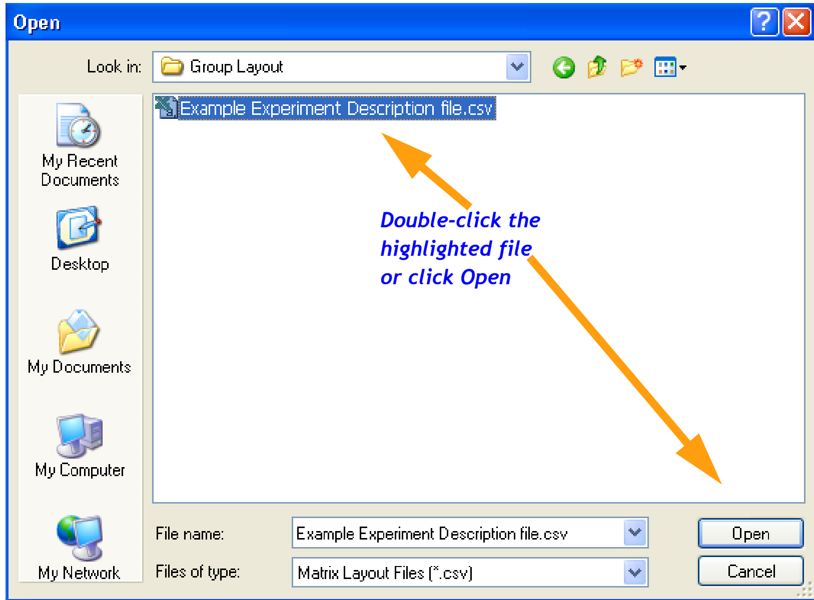


Figure 2-23 Open Pop-Up

3. Samples on the selected SAM will be mapped into groups according to the Group Layout file you applied; those groups will be displayed in the **Groups** listbox (Figure 2-24).

NOTE:

See *Set Up & Apply Group Layout Files* on page 2-25 for a file format template example.

4. To display the samples for a group in the **Groups** listbox, click the plus sign (+) next to that group (Figure 2-24).

Figure 2-24 Groups Displayed per Applied Group Layout File

NOTE:

Each Beadchip layout must be saved in a separate group layout file.

Set Up & Apply Sample Sheet

To help BeadStudio effectively track your samples and assay, Illumina recommends you create a Sample Sheet.

Create your Sample Sheet in MS Excel, according to the guidelines provided in Table 2-1.

NOTE:

The Sample Sheet is used in Illumina's Genotyping products as well as its Gene Expression products. Therefore, some columns in the Sample Sheet are relevant to one or the other, but not both, product lines). Table 2-1 displays the Sample Sheet columns that BeadStudio uses in **BLUE**.

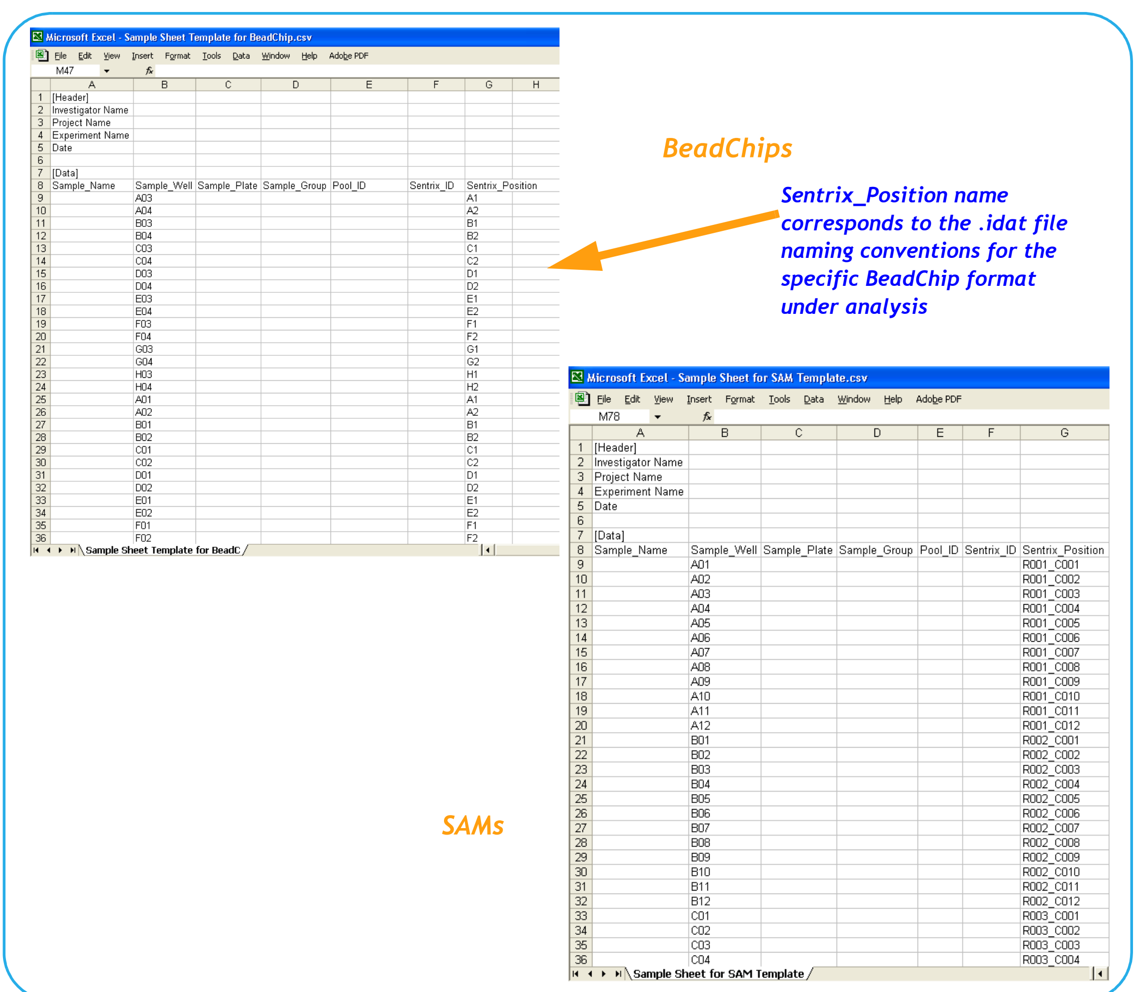
Table 2-1 Sample Sheet Guidelines

	Description	*Optional (O) or Required (R)*
Sample_Name	For example, S12345. If not user-specified, the BeadStudio application will assign a default sample name, concatenating the sample plate and sample well names.	O
Sample_Well	For example, A01. The well containing the specific sample in the 96-well sample plate.	O
Sample_Plate	For example, XXXXXXXXX-RNA. User-specified name for the plate containing RNA samples.	O
Sample_Group	For example, Group_1 User-specified name of the sample group. Note: If Sample_Group is missing, BeadStudio creates one group with the name "Default Group"	R
Pool_ID	Not Used for Direct Hyb.	
	DASL™ Assay For example, GS0006187-DAP. Name of the DAP.	R

Table 2-1 Sample Sheet Guidelines (continued)

	Description	Optional (O) or Required (R)
Sentrix_ID	For example, 1167988 SAM or BeadChip ID	**R**
Sentrix_Position	For example, R001_C001 for a SAM, A1 for a BeadChip. For SAMs, the SAM sample to which the sample is hybridized. For BeadChips, the section to which the sample is hybridized.	**R**
NOTES	• Your sample sheet header may contain any, and as much, information as you choose. • Your sample sheet may contain any number of columns you choose. • Your sample sheet must be in a comma-delimited (.csv) file format	

Figure 2-25 provides examples of the flexible Sample Sheet format for both BeadChip and SAM. The BeadStudio Documentation CD (Illumina part # 11181791) includes electronic, read-only Sample Sheet template files (*Sample Sheet Template for BeadChip.csv* and *Sample Sheet Template for SAM.csv*) you can copy and use.



	A	B	C	D	E	F	G	H
1	[Header]							
2	Investigator Name							
3	Project Name							
4	Experiment Name							
5	Date							
6								
7	[Data]							
8	Sample_Name	Sample_Well	Sample_Plate	Sample_Group	Pool_ID	Sentrix_ID	Sentrix_Position	
9		A03					A1	
10		A04					A2	
11		B03					B1	
12		B04					B2	
13		C03					C1	
14		C04					C2	
15		D03					D1	
16		D04					D2	
17		E03					E1	
18		E04					E2	
19		F03					F1	
20		F04					F2	
21		G03					G1	
22		G04					G2	
23		H03					H1	
24		H04					H2	
25		A01					A1	
26		A02					A2	
27		B01					B1	
28		B02					B2	
29		C01					C1	
30		C02					C2	
31		D01					D1	
32		D02					D2	
33		E01					E1	
34		E02					E2	
35		F01					F1	
36		F02					F2	

	A	B	C	D	E	F	G
1	[Header]						
2	Investigator Name						
3	Project Name						
4	Experiment Name						
5	Date						
6							
7	[Data]						
8	Sample_Name	Sample_Well	Sample_Plate	Sample_Group	Pool_ID	Sentrix_ID	Sentrix_Position
9		A01					R001_C001
10		A02					R001_C002
11		A03					R001_C003
12		A04					R001_C004
13		A05					R001_C005
14		A06					R001_C006
15		A07					R001_C007
16		A08					R001_C008
17		A09					R001_C009
18		A10					R001_C010
19		A11					R001_C011
20		A12					R001_C012
21		B01					R002_C001
22		B02					R002_C002
23		B03					R002_C003
24		B04					R002_C004
25		B05					R002_C005
26		B06					R002_C006
27		B07					R002_C007
28		B08					R002_C008
29		B09					R002_C009
30		B10					R002_C010
31		B11					R002_C011
32		B12					R002_C012
33		C01					R003_C001
34		C02					R003_C002
35		C03					R003_C003
36		C04					R003_C004

Figure 2-25 Sample Sheet Examples

1. From the **Groups** pane toolbar, click **Load from sample sheet...** .

2. When the **Open Sample Sheet** pop-up appears, navigate to the appropriate file and click **Open** (Figure 2-26).

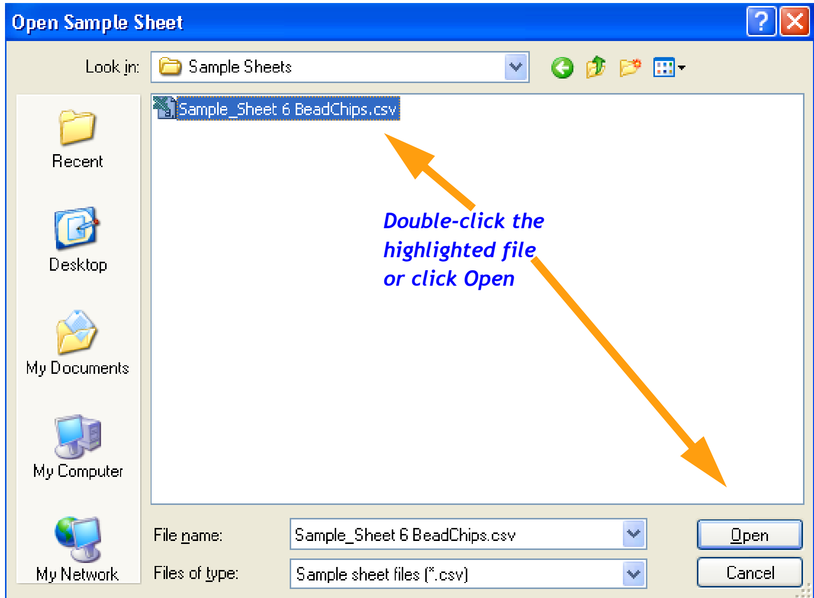


Figure 2-26 Open Sample Sheet Pop-Up

NOTE:

If data files are missing, a warning message will alert you (Figure 2-27). If this message pops up: 1) check your Sample Sheet for errors; 2) ensure that your data files are in the proper location; and 3) ensure that the repository is correctly selected.

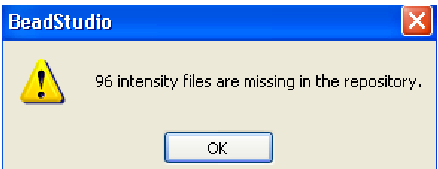


Figure 2-27 Missing Sample Sheet Data Files Warning Message

Gene Analysis Output Files

Following gene analysis, the files listed below are created and placed in the directory for the associated experiment:

- XXXXXX_gene_profile.csv ***OR*** XXXXXX_gene_profile.txt
- XXXXXX_gene_probe_profile.csv ***OR*** XXXXXX_gene_probe_profile.txt
- XXXXXX_qcinfo.csv ***OR*** XXXXXX_qcinfo.txt

  Where "XXXXXX" is the experiment name.
- If you have selected the MAGE-ML output format, the file created will be: XXXXXX_MAGE-ML.xml

Table 2-2 provides brief descriptions of the output files.

Table 2-2 Output File Descriptions

Output filename	Description
XXXXXX_gene_profile	Intensity data and various quality scores reported at the gene level. Signals from probes for the same gene are combined to give a single value for the gene.
XXXXXX_gene_probe_profile	Intensity data and various quality scores reported at the probe level. Each probe is listed individually. This file is only generated for experiments using more than one probe per gene.
XXXXXX_gene_diff	Intensity determining if gene expression levels have changed between two experimental groups.
XXXXXX_qcinfo	Intensity data for categories of experimental control probes. The categories of controls reported in the file are described in the System Manual *System Controls* appendix for your specific product.
NOTE: Detailed descriptions of the contents of these files is provided in Tables 2-3, 2-5, and Table 2-5.	

Refer to Figures 2-28 through 2-31 and Tables 2-3 and 2-5 for example output files and column heading descriptions, respectively.

Microsoft Excel - Experiment 1_gene_profile.csv

File Edit View Insert Format Tools Data Window Help Adobe PDF — Type a question for help

Y34 fx 0.01036478

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Illumina Inc. BeadStudio X version 1.4.0.0															
2	Normalization = none															
3	Array Content = RefSeq_Content.xml															
4	Error Model = none															
5	DateTime = 10/12/2004 3:03 PM															
6	Local Settings = en-US															
7																
8	TargetID	MIN_Signa	AVG_Sign	MAX_Sign	NARRAYS	ARRAY_S	BEAD_STI	Avg_NBEA	Detection-(	MIN_Signa	AVG_Sign	MAX_Sign	NARRAYS	ARRAY_S	BEAD_STI	Avg
9	GI_100470	330	332.3	334.6	2	3.3	22.6	29	0.31104	274.9	308	324.7	3	28.7	20.1	
10	GI_100470	849.3	888.4	927.6	2	55.3	43.2	26	1	700.6	834.6	917.8	3	117.2	42.8	
11	GI_100470	1747.1	1975.2	2203.2	2	322.5	100.5	36	1	1686.5	1811.9	1922.9	3	118.8	84.3	
12	GI_100470	4258.4	4342.9	4427.5	2	119.5	144.1	50	1	4198.8	4772.2	5218	3	521.5	173	
13	GI_100471	5732.5	6606.8	7481.2	2	1236.5	350.7	31	1	5160.6	5711.3	6137.7	3	500.3	288.8	
14	GI_100471	2012.2	2313.3	2614.4	2	425.8	124.9	32	1	2058	2403.2	2679.4	3	316.4	118.1	
15	GI_100471	279.3	287.2	295.2	2	11.2	17.4	44	0.051778	230.1	253.1	272.7	3	21.5	12.6	
16	GI_100471	641.6	697.4	753.3	2	79	24.1	45	1	626.7	652.7	697.4	3	38.9	20.6	
17	GI_100471	532.3	556.7	581.1	2	34.5	28.4	33	1	500.7	515.8	541.6	3	22.5	28.8	
18	GI_100484	244.2	247.6	251	2	4.8	10.3	35	0.004113	225	237.2	252.3	3	13.9	16.5	
19	GI_100925	318.5	336.9	355.4	2	26.1	14.7	44	0.36351	329.9	336.4	345	3	7.7	16.4	
20	GI_100925	540.8	587	633.2	2	65.3	34.7	33	1	484.6	540.5	619.8	3	70.6	30.4	
21	GI_100925	619.3	682.7	746.1	2	89.7	28.6	47	1	618.9	661.5	741.6	3	69.4	28.1	
22	GI_100926	415.4	416.9	418.5	2	2.2	28.3	30	0.951802	365.4	407.6	451.7	3	43.2	27	
23	GI_100926	659.6	690.9	722.2	2	44.3	48.9	23	1	600.8	647.1	706.2	3	53.9	36.5	
24	GI_100926	618.8	660.3	701.8	2	58.7	27.5	42	1	646.6	663.6	683.2	3	18.5	30.2	
25	GI_100926	378.3	398.6	418.9	2	28.7	21.9	31	0.87944	342.9	387.8	431.3	3	44.3	15.1	
26	GI_100926	3013.8	3708	4402.2	2	981.8	115.8	42	1	2987.8	3318.2	3503.9	3	286.9	116.7	
27	GI_100926	472	486.5	501	2	20.5	31.9	32	0.999684	391.2	478.6	532.3	3	76.4	30.4	
28	GI_100926	323.9	355.9	388	2	45.3	20.4	36	0.528093	286.3	306.7	334.7	3	25.1	22	
29	GI_100926	885.3	933.8	982.3	2	68.6	35.2	45	1	775	800.8	833.1	3	29.6	25.6	
30	GI_100926	380	387.2	394.5	2	10.3	18.6	27	0.815776	360.2	387.3	439.8	3	45.5	20.1	
31	GI_100926	1631.3	1829.4	2027.5	2	280.2	70.5	42	1	1525.6	1541.3	1566.8	3	22.3	60.6	
32	GI_100926	840.2	903.7	967.2	2	89.8	48.8	28	1	737	798.4	863.5	3	63.3	44	
33	GI_100926	345.3	366.1	386.8	2	29.3	19	46	0.65401	349.5	387.1	435.7	3	44.1	26.1	
34	GI_101906	251.2	279.4	307.7	2	40	16.3	29	0.035913	214.4	239.5	258.1	3	22.6	14.3	
35	GI_101906	658.2	669.6	681.1	2	16.2	38.7	31	1	675.6	692.4	711.9	3	18.3	42	
36	GI_101906	420.5	432.9	445.4	2	17.6	20.7	24	0.98161	425	441.9	451.4	3	14.7	24.3	
37	GI_101906	352.9	354.8	356.7	2	2.6	19.9	40	0.531922	291.1	307.3	317.8	3	14.2	16.1	

Experiment 1_gene_profile

Figure 2-28 Example XXXXXX_gene_profile File

NOTE:

The example shown in Figure 2-28 was generated in a “column-oriented” format. When generated in “column-oriented” format, the GroupID column that appears in the “row-oriented” format (not shown) disappears, and the GroupID is appended to all column headings according to the group with which they are associated.

Table 2-3 gene_profile Column Descriptions

Column Heading	Description
GroupID	Identifies the experimental group.
TargetID	Identifies the gene name
ProbeID	Identifies the bead type
MIN Signal *	Minimum intensity of the Bead Type/target in the group
AVG Signal *	Average intensity of the Bead Type/target in the group
MAX Signal *	Maximum intensity of the Bead Type/target in the group
NARRAYS	Number of samples in the group
ARRAY_STDEV	Standard deviation associated with sample-to-sample variability within the group (undefined when the group contains a single sample)
BEAD_STDEV	Average standard deviation associated with bead-to-bead variability for the samples in the group
AVG_NBEADS	Average number of beads per bead type representing probes for the gene
Detection	1-p-value computed from the background model characterizing the chance that the target sequence signal was distinguishable from the negative controls
DASL™ Assay 0/1	Used for masking unverified probes in DASL mode.

In DASL Mode, signal is the sum of the red and green channel signals.
Bead stdev is an average of bead stdev on red and green channels.

* *In the case of groups containing only one sample, MIN, AVG, and MAX signals are equal.*

Microsoft Excel - Experiment 1_gene_probe_profile.csv

File Edit View Insert Format Tools Data Window Help Adobe PDF

AA37

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Illumina Inc. BeadStudio X version 1.4.0.0															
2	Normalization = none															
3	Array Content = HumanTox1.xml															
4	Error Model = none															
5	DateTime = 10/12/2004 3:12 PM															
6	Local Settings = en-US															
7																
8	TargetID	ProbeID	MIN_Signa	AVG_Sign	MAX_Sign	NARRAYS	ARRAY_S	BEAD_STI	Avg_NBEA	Detection-(	MIN_Signa	AVG_Sign	MAX_Sign	NARRAYS	ARRAY_S	BE
9	GI_100926	80	98.3	141.8	308.9	60	42.1	6.2	27	0.282158	98.3	137.5	308.9	36	39.9	
10	GI_100926	752	448.5	2549.2	4217.7	60	932.4	99.9	30	0.999419	960.2	2474.2	4044.7	36	701.1	
11	GI_107165	61	190	456.8	1575.5	60	204.6	19.6	28	0.46694	210.6	453.1	1575.5	36	224.7	
12	GI_107165	4303	102.7	450.7	738.9	60	144.6	21.2	30	0.509246	102.7	478.3	738.9	36	139.5	
13	GI_108349	101	106.9	215.4	1262.6	60	168.7	11.6	28	0.307751	106.9	182.1	451	36	90.6	
14	GI_108349	2811	99.6	136.5	1070.7	60	125.7	11.2	27	0.279279	99.6	125.7	293.2	36	34.2	
15	GI_108349	2831	133.4	1650.9	2786.1	60	722.4	81.3	30	0.990133	643	1815.1	2588.5	36	691.8	
16	GI_108349	5070	1691.5	2470.8	3317.7	60	398.1	108.6	26	0.999334	1692.3	2399.1	2969.8	36	330	
17	GI_108350	273	395.2	1506.1	2738.7	60	457.8	65.1	29	0.959617	395.2	1464.9	2347.3	36	395.6	
18	GI_108350	4969	869.1	2684.6	3417.6	60	409.4	105.4	27	0.999906	1706.5	2710.6	3146.3	36	280.7	
19	GI_108350	230	162.4	896.4	1290.9	60	226.3	40.3	27	0.779958	542.4	934.8	1290.9	36	160.5	
20	GI_108350	2833	1170.7	2436.4	3473.7	60	664.2	97.6	30	0.999861	1503.7	2605.4	3351.1	36	554.3	
21	GI_108351	153	127.4	761.5	1560.6	60	364.7	35.6	28	0.673738	341.7	769.3	1560.6	36	349.2	
22	GI_108351	1331	115	933	3247.8	60	1009.5	46.5	33	0.586859	116	651.2	2478.8	36	791.3	
23	GI_108351	54	547.9	1876.3	2789.2	60	436.4	75	30	0.990404	1232.9	1884.7	2446.4	36	330.5	
24	GI_108351	4115	104.3	468.6	1313	60	335	27.1	27	0.534934	136.5	530	1313	36	341.5	
25	GI_108351	56	163.4	1516.3	2395.9	60	486.5	66.1	30	0.957998	829.5	1574.8	2395.9	36	396.6	
26	GI_108351	74	174.2	1177.5	3010.2	60	437.7	47.2	29	0.863234	535.5	1119.6	1920.4	36	295.5	
27	GI_108351	1022	235.2	1894.5	3898.5	60	634	85.6	30	0.990149	598.6	1946.6	3898.5	36	629.4	
28	GI_108351	75	158.3	1815.9	3136.7	60	604.9	88.3	27	0.984104	812	1763	2665.4	36	399.2	
29	GI_108351	1842	234.5	1141.3	2081.1	60	400.2	54.7	30	0.843986	741.1	1118.6	1962.9	36	300.9	
30	GI_108351	905	332.6	2255	3676.2	60	970.3	88.7	30	0.999651	436.4	2455.9	3676.2	36	931.3	
31	GI_108351	4857	146.6	547.4	1379.8	60	273.9	25.1	29	0.504853	213.9	492.8	1176.3	36	213.4	
32	GI_108351	1650	117.3	183.3	305.4	60	42.1	7.4	31	0.313885	121.4	184.8	262.7	36	36.6	

Figure 2-29 Example XXXXXX_gene_probe_profile File

NOTE:

The example shown in Figure 2-29 was generated in a "column-oriented" format. When generated in "column-oriented" format, the GroupID column that appears in the "row-oriented" format disappears, and the GroupID is appended to all column headings according to the group with which they are associated.

The output file column headings are essentially the same for gene_probe_profile file as for the gene_profile file. See Table 2-3.

Microsoft Excel - Brain vs. Liver Differential Expression - April 27, 2005_gene_diff.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Illumina Inc. BeadStudio version 1.5.0.32																		
2	Normalization = rank invariant																		
3	Array Cont	REFSEQ2	BETA_PN11193506A_11193418.xml																
4	Error Model = Illumina custom																		
5	DateTime = 3/14/2005 5:18 PM																		
6	Local Settings = en-US																		
7																			
8	TargetID	MIN_Signa	AVG_Sign	MAX_Sign	NARRAYS	ARRAY_S	BEAD_STI	Avg_NBEA	Detection-(	MIN_Signa	AVG_Sign	MAX_Sign	NARRAYS	ARRAY_S	BEAD_STI	Avg_NBEA	Detection-(	DiffScore-(	Concordance-
9	GI_100470	-217.221	-58.4382	-2.14412	4	105.891	55.3711	23	0.200257	-16.4037	2.26799	15.8675	4	14.0107	3.87035	28	0.758665	4.525339	1
10	GI_100470	12.7944	777.225	2240.01	4	998.167	77.7199	26	1	12.7637	3661.75	5560.47	4	2499.32	123.865	28	1	14.93077	1
11	GI_100470	14.5858	82.2508	125.211	4	47.3832	88.1703	21	0.997433	67.2704	82.5376	107.484	4	18.0455	8.29963	21	0.998716	0.012442	1
12	GI_100470	142.521	199.299	283.069	4	68.8922	78.7155	24	1	144.167	321.693	397.15	4	119.951	16.9637	31	1	5.30527	1
13	GI_100470	51.4251	86.1386	179.453	4	62.2975	79.3211	26	0.993582	28.5322	45.4389	84.3237	4	26.0934	5.11369	25	0.994865	-2.12	1
14	GI_100471	482.589	542.49	651.762	4	75.0719	168.01	29	1	530.146	552.742	579.455	4	22.1244	20.2271	29	1	0.198995	1
15	GI_100471	-7.38113	5.20672	23.5638	4	13.07	117.598	16	0.509628	-3.07645	4.37764	15.9906	4	8.42779	4.92885	18	0.750963	-0.04388	1
16	GI_100471	-163.961	1302.61	2155.24	4	1027.74	132.743	16	1	537.852	1070.77	2236.67	4	789.969	61.3267	17	1	-1.42059	1
17	GI_100471	47.6557	73.1958	103.352	4	24.5757	82.5027	29	0.994865	6.12585	40.3528	101.932	4	42.3196	5.72642	34	0.992298	-1.66271	1
18	GI_100471	-82.5197	-25.0353	1.39281	4	38.8066	52.7497	18	0.269576	-6.733	-2.55045	0.290878	4	3.01447	3.88768	28	0.417202	1.477926	1
19	GI_100471	104.725	149.264	187.203	4	36.9451	121.806	23	1	21.2417	77.8953	198.706	4	83.011	6.55031	31	0.997433	-3.22911	1
20	GI_100471	-449.657	-86.8451	81.5676	4	244.197	57.6089	22	0.97561	-3.42268	149.16	228.579	4	106.547	10.3562	26	1	10.54499	1
21	GI_100471	-2.53302	91.4028	362.574	4	180.803	66.8398	33	0.727856	-6.61327	1.09663	5.90508	4	5.38096	3.4459	28	0.604621	-4.47858	1
22	GI_100484	6.24606	55.8056	198.33	4	95.0275	79.4255	28	0.856226	-3.43491	2.03055	7.55036	4	5.46115	3.68953	28	0.684211	-3.28634	1
23	GI_100925	14.4012	5968.95	9390.01	4	4330.28	240.11	23	1	4.12108	3603.98	11119.5	4	5161.64	107.236	26	1	-3.16334	1
24	GI_100925	16.4421	172.111	625.639	4	302.399	125.391	22	0.971759	17.0193	34.7831	46.193	4	13.9667	5.10663	26	0.992298	-4.21825	1
25	GI_100925	-96.852	173.607	303.404	4	182.543	74.1822	33	1	224.793	243.501	261.435	4	15.2929	12.0526	26	1	2.976903	1
26	GI_100926	-2.46318	82.2599	323.016	4	160.547	156.139	25	0.736842	-7.37297	-1.06679	5.61948	4	6.35206	3.58988	27	0.498074	-4.58967	1
27	GI_100926	-2.54076	66.559	258.229	4	127.977	99.0858	34	0.722721	0.782959	7.34418	12.4992	4	5.08244	4.42133	30	0.856226	-3.59413	1
28	GI_100926	-480.897	-86.5464	50.032	4	262.926	60.0903	17	0.98973	38.7138	45.6048	63.5908	4	12.0095	5.81132	22	0.994865	4.75288	1
29	GI_100926	-317.893	-51.0718	57.4594	4	178.518	89.9687	16	0.983312	26.7981	94.4003	134.952	4	49.191	9.02925	18	1	8.371561	1
30	GI_100926	-138.725	571.323	966.441	4	490.001	58.0594	19	1	333.227	529.503	1011.84	4	322.97	25.2252	27	1	-0.51791	1
31	GI_100926	-62.2099	2.39081	30.236	4	43.3128	108.911	21	0.955071	15.4333	41.8312	54.4369	4	18.1022	4.65938	27	0.994865	2.603811	1
32	GI_100926	-34.1907	-5.6319	5.23896	4	19.0672	56.3735	25	0.673941	-3.96516	3.01955	10.5185	4	6.95153	4.14678	24	0.684211	0.494216	1
33	GI_100926	-453.767	-94.8622	30.8878	4	239.309	95.317	19	0.946085	27.2234	44.4041	58.1475	4	12.8431	7.15918	21	0.994865	5.720199	1
34	GI_100926	-1.76211	94.8973	370.88	4	184.029	85.4258	24	0.783055	-1.68697	5.29805	10.2194	4	5.82907	3.94061	35	0.816431	-4.3456	1
35	GI_100926	105.171	184.479	242.057	4	58.4807	51.499	22	1	183.231	195.714	213.167	4	13.9972	10.0616	27	1	0.400796	1
36	GI_100926	24.0892	148.841	478.457	4	220.058	124.581	22	0.991014	18.1103	46.6569	61.2336	4	19.5609	5.22441	23	0.994865	-4.18833	1
37	GI_100926	-119.898	-6.63977	47.0997	4	76.3478	76.404	20	0.970475	7.16556	14.6579	25.5995	4	7.80164	4.8152	24	0.938383	1.314714	1
38	GI_101906	-10.0417	56.5275	243.039	4	124.41	88.9303	23	0.405648	-8.41698	-1.56848	4.37644	4	5.30472	4.17528	22	0.442875	-3.61494	1
39	GI_101906	-6.25144	51.5251	195.611	4	96.314	100.878	12	0.735558	-3.92133	3.51727	10.4189	4	5.99244	6.67606	13	0.713736	-2.89008	1
40	GI_101906	-0.57584	81.4139	323.441	4	161.355	142.273	15	0.70475	-6.63161	-1.91896	3.17917	4	4.2789	4.81849	20	0.467266	-4.5674	1
41	GI_101906	-8.24683	23.9033	94.2822	4	47.3566	150.503	14	0.617458	-6.57349	1.6818	12.2211	4	9.65064	5.41567	17	0.681643	-1.26534	1
42	GI_101906	-97.4921	175.861	324.413	4	196.192	71.9899	27	1	-13.1307	100.677	344.34	4	166.114	6.06631	23	0.998716	-2.40679	1
43	GI_101906	26.8952	114.932	289.022	4	118.658	92.4012	19	0.993582	30.0152	46.2171	62.9207	4	13.4398	6.71875	28	0.994865	-3.77025	1
44	GI_101906	7.4438	21.4284	50.6308	4	19.6995	66.8638	25	0.8819	7.69662	11.9654	16.2436	4	4.28759	5.08599	27	0.915276	-0.5056	1
45	GI_101906	15.5967	49.7644	136.469	4	57.9236	80.3437	25	0.95122	13.7383	18.0469	23.1508	4	3.87741	4.3863	30	0.961489	-1.7686	1
46	GI_101906	-235.017	-69.015	-12.6803	4	110.671	59.8329	27	0.032092	-9.85629	-5.92684	3.24063	4	6.14306	3.15098	25	0.181001	4.605869	1
47	GI_101906	-168.464	-39.0651	11.8149	4	86.4402	131.468	10	0.626444	4.81615	12.2968	19.2326	4	5.92992	6.84515	17	0.929397	4.182477	1
48	GI_101906	-14.2863	96.9001	399.522	4	201.883	112.733	11	0.585366	1.22207	2.641	6.57169	4	2.62171	5.92455	17	0.676508	-4.18242	1
49	GI_101906	0.926102	32.897	120.741	4	58.5934	82.7568	19	0.715019	-3.50909	-0.63568	3.25723	4	2.92368	3.73412	22	0.501926	-1.98184	1

Brain vs. Liver Differential Ex

Figure 2-30 Example XXXXXX_gene_diff File

Table 2-4 gene_diff Column Descriptions

Column Heading	Description
TargetID	Identifies the experimental group.
MIN Signal *	Minimum intensity of the Bead Type/target in the group
AVG Signal *	Average intensity of the Bead Type/target in the group
MAX Signal *	Maximum intensity of the Bead Type/target in the group
NARRAYS	Number of samples in the group
ARRAY_STDEV	Standard deviation associated with sample-to-sample variability within the group (undefined when the group contains a single sample)
BEAD_STDEV	Average standard deviation associated with bead-to-bead variability for the samples in the group
AVG_NBEADS	Average number of beads per bead type representing probes for the gene
Detection	1-p-value computed from the background model characterizing the chance that the target sequence signal was distinguishable from the negative controls
DiffScore	Statistical confidence that the gene's expression has changed with respect to a reference group
Concordance	Number of probes showing upregulated signal compared to the number of probes showing downregulated signal
DASL™ Assay 0/1	Used for masking unverified probes in DASL mode.

In DASL Mode, signal is the sum of the red and green channel signals. Bead stdev is an average of bead stdev on red and green channels.

* *In the case of groups containing only one sample, MIN, AVG, and MAX signals are equal.*

Microsoft Excel - Experiment 1_qcinfo.csv

File Edit View Insert Format Tools Data Window Help Adobe PDF

AH31 fx 0.50000002

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Illumina Inc. BeadStudio X version 1.4.0.0														
2	Normalization = none														
3	Array Content = HumanTox1.xml														
4	Error Model = none														
5	DateTime = 10/12/2004 3:12 PM														
6	Local Settings = en-US														
7															
8	ArrayID	AVG-Signa	AVG-SeqV	AVG-Dete	AVG-Signa	AVG-SeqV	AVG-Dete	AVG-Signa	AVG-SeqV	AVG-Dete	AVG-Signa	AVG-SeqV	AVG-Dete	AVG-Signa	AVG-SeqV
9	1202085_D	110.2	3.6	0.166843	121.5	8.7	0.176552	3331.1	745.5	1	3445.2	23.4	1	1131	697.9
10	1202085_D	96.5	2.4	0.151658	117	10.5	0.165394	4477.8	442.8	1	4532.2	815.2	1	1423.2	897.8
11	1202085_D	105.2	8.2	0.18247	115.6	13.5	0.190978	4000	917.9	1	4517.4	31.5	1	1387.2	876.1
12	1202085_D	116.6	1.8	0.206321	116.4	7.9	0.206205	4384.7	697.9	1	4410.1	122.6	1	1338.4	855.3
13	1202085_D	102.5	9.5	0.203823	126	18	0.224079	4629.5	873.8	1	4468.6	158.8	1	1329.8	862.5
14	1202085_D	109.7	2.5	0.224982	113.3	6.6	0.228234	4524.8	830.3	1	4813.8	350.3	1	1296.8	867.7
15	1202085_D	113.9	7.5	0.243721	112.6	5.8	0.242434	4792.5	587.5	1	4776.3	24.4	1	1294.7	894.5
16	1202085_D	106.1	9.2	0.26366	127.1	13.9	0.284674	4687	858.5	1	4513	53.5	1	1155.1	807.2
17	1202085_D	195.7	32.9	0.291621	199.2	16	0.294973	4490.9	1053	1	4615	152.4	1	1285.3	890.2
18	1202085_D	100	1	0.201119	103.7	12.9	0.202359	4274.1	854.9	0.999988	4368	40.8	0.999993	1413.5	842.9
19	1202085_D	103.2	6.1	0.204135	108.6	15.3	0.205846	4983.2	746.1	0.999999	4817.6	225.4	0.999996	1597.4	928.6
20	1202085_D	100.7	2.6	0.197923	114.2	9.2	0.202408	4150.9	650.9	0.99996	4265.7	92.8	0.999977	1437.5	838
21	1202085_E	101.4	10.5	0.181751	109.2	14.4	0.187061	4789.3	908.7	1	4913.2	74.8	1	1455	973.2
22	1202085_E	108.9	9.7	0.163856	122.2	16.6	0.17319	4292.2	981.1	1	3760.7	122.6	1	1391.8	859.8
23	1202085_E	115.2	2.4	0.177669	126.9	17.5	0.185363	4502.8	638.3	1	4700.6	177.1	1	1496.4	948.2
24	1202085_E	101.1	1.6	0.197996	126.6	15.2	0.216543	4606.3	1013.1	1	4734.9	66.1	1	1441.8	913.6
25	1202085_E	110.6	4.9	0.201248	121.1	5.5	0.209759	4461.3	719.8	1	4495.4	170.7	1	1338.9	869.2
26	1202085_E	115.9	8.3	0.233449	115	11.3	0.232679	4547.9	773.9	1	4244.7	128	1	1285.3	853.9
27	1202085_E	111.2	12.4	0.247984	123.1	11.2	0.257667	4811.5	812	1	4787.8	157.7	1	1323	901.1
28	1202085_E	100.4	2	0.260493	106.6	5.4	0.266898	4314.8	911.2	1	4493.3	156.9	1	1128	795.7
29	1202085_E	99.8	1.2	0.295004	112.5	8.8	0.311743	4072.4	837.4	1	3992	388.6	1	976.7	731.3
30	1202085_E	108.6	11.9	0.199828	110.9	10.6	0.200528	4645.1	722.8	0.99998	4678.5	199.5	0.999983	1485.2	867.7
31	1202085_E	118.2	8.5	0.214064	118	0.9	0.213994	4692.3	635.2	0.999994	4626.8	110.5	0.999992	1461.4	848.3

Figure 2-31 Example XXXXXX_qcinfo File for Group Gene Analysis

NOTE:

The qcinfo file contains statistics for various controls and allows you to access performance measures for the experiment. It contains raw data, not normalized data. The categories of controls reported in the file are described in the System Manual *System Controls* appendix for your specific product.

Table 2-5 qcinfo Column Descriptions

<table>
<tr><th>Column Heading</th><th>Description</th></tr>
<tr><td>ArrayID</td><td>Uniquely identifies the sample by combined BeadChip ID and position of the sample on the SAM or BeadChip.</td></tr>
<tr><td>All other column headings</td><td>Names of control categories for which BeadStudio computes performance metrics.
For each control category:
- AVG-Signal-XXX
▹ average signal of probes of a particular type (for example, probes corresponding to Cy3_hyb_low control targets)
- AVG-SeqVAR-XXX
▹ average standard deviation of probes corresponding to the same target transcript, where averaging occurs for all targets of the same type (for example, all genes, all housekeeping controls)
- AVG-Detection-XXX
▹ average detection 1-p-value for probes of a particular type
NOTE:
“XXX” is a probe type name. Types are described in the “System Controls” appendix in each Illumina product manual.</td></tr>
</table>

Creating the Mask File

If genomic DNA was used in a DASL Assay to verify probe performance, you can select probes that should be excluded from further analysis. Because all probes are designed to be intraexonic, all probes should be detectable when genomic DNA is used as a sample. Therefore, the Detection score reported in the *XXXXXX_gene_probe_profile* file can be used as an objective measure of probe performance on genomic DNA.

Illumina recommends excluding probes that have a Detection score of less than 1 on genomic DNA. However, you may define your own exclusion criteria.

In DASL mode, the *XXXXXX_gene_probe_profile* file contains an additional column, with the heading **0/1**. By default, this column will display a 1 in all rows containing probe-level data.

To exclude a probe:

1. Change the 1 to 0 in the **0/1** column.
2. Delete all the data columns between the **Probe_ID** column and the **0/1** column.
3. Save the file as a *.csv file in the same repository where the Content Descriptor file is stored. The file need not conform to a naming convention.

NOTE:

Any *.csv file present in the same repository as Content Descriptor files will appear in the Experiment Parameters pulldown menu. To avoid confusion, Illumina advises the use of separate repositories for Content Descriptor and for SAM/BeadChip data.

Browse History

BeadStudio maintains a historical record of repository browse actions. To view your browse history, click the down arrow to the left of the Browse button (Figure 2-32).

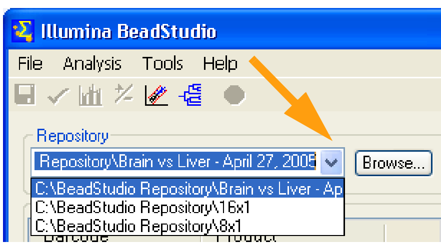


Figure 2-32 Browse History Pulldown List

Clear or Copy Message Log Feature

To either copy the contents of the message log pane or to clear the message log pane, right-click anywhere in the pane, then click either **Copy** or **Clear** (Figure 2-33).

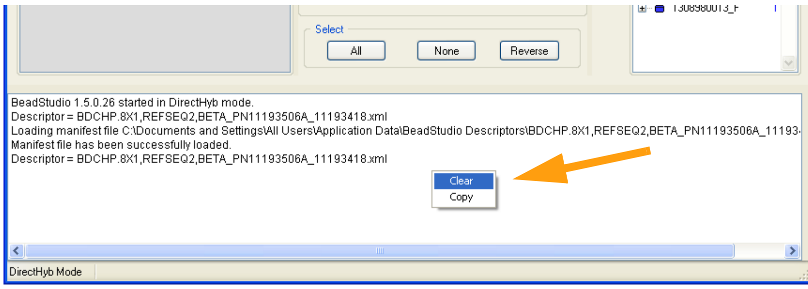


Figure 2-33 Message Pane Clear/Copy Button

NOTES

Data Visualization

Topics

- Introduction *3-2*
  - *Accessing the Data Visualization Tools 3-3*
    - *From the New Experiment Wizard 3-3*
    - *From the BeadStudio Main Page 3-4*
- Scatter Plots *3-7*
  - *Scatter Plot Context Menu Functions 3-16*
  - *Finding Genes in the Scatter Plot 3-19*
  - *Other Scatter Plot Functionalities 3-27*
- Cluster Analysis *3-28*
  - *Analyze Clusters 3-30*
  - *Dendrogram Context Menu Selections 3-34*
  - *View the Sub-Tree List Directly in the Dendrogram 3-35*
- Copy/Paste Clusters *3-36*
  - *From Scatter Plot to Dendrogram 3-36*
  - *From Dendrogram to Scatter Plot 3-38*
- Control Summary Reports *3-40*
  - *Control Summary Reports for DirectHyb 3-40*
  - *Control Summary Reports for the DASL™ Assay 3-43*

Introduction

This chapter describes BeadStudio's data visualization functions, used to create:

- Scatter plots (using **Start scatter plot** tool)
- Cluster analysis diagrams (dendrograms) (using **Start cluster analysis** tool)
- Control Summary Reports (using the **Tools | Visualize Existing Data...** menu bar selection)

Use these tools to explore the data you created using the Gene Analysis or Differential Expression Analysis tools (described in Chapter 2, *Experiment Creation & Analysis*).

Accessing the Data Visualization Tools

Access the data visualization functions either from the **New Experiment Wizard** (see below) or from the **BeadStudio main page** (page 3-4).

From the New Experiment Wizard

In the last dialog box displayed in the **New Experiment Wizard** (**How do you want to view your data?** pane -- Figure 3-1):

1. Select the radio button next to the desired tool.
2. Click **OK**.

The wizard will automatically launch the selected tool using the experimental data most recently processed.

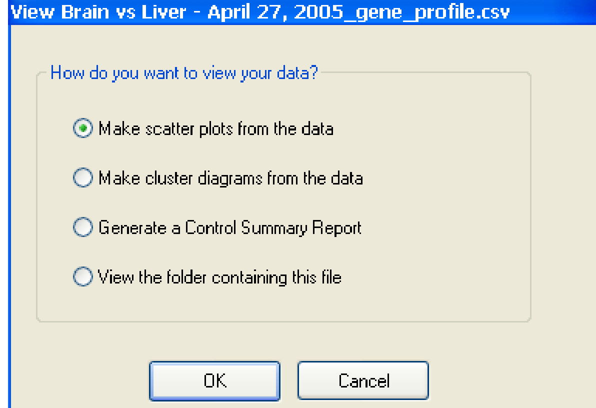


Figure 3-1 How do you want to view your data? Dialog Box

From the BeadStudio Main Page

You can use the main page menu bar or tool bar to visualize your data. These two methods are described below.

1. At the menu bar:
   a. Select **Tools | Visualize Existing Data...** (Figure 3-2).

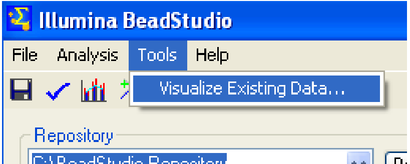


Figure 3-2 Tools Menu

   b. At the **Please select the data file you want to visualize** box, browse to your processed data output files, highlight the file you want to use (gene_profile, gene_probe_profile, or gene_diff), and click **Open** (Figure 3-3).

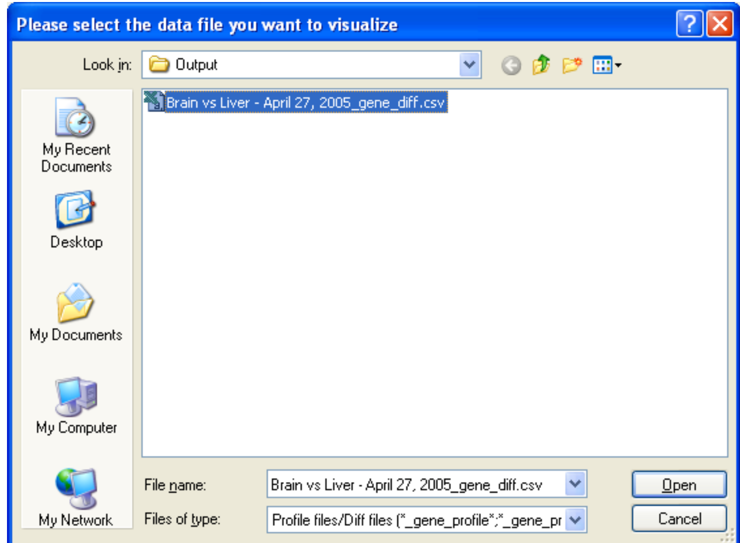


Figure 3-3 Browse to the Data File

NOTE:

Only files generated in column-oriented format are supported.

c. At the next dialog box, **How do you want to view your data?** pane (same as shown in Figure 3-1), select the radio button next to the desired tool and click **OK**.

d. BeadStudio will launch the selected tool using the experimental data you specified in step b, above.

2. As an alternate method, from the tool bar:

a. Click **Start scatter plot** or **Start cluster analysis** (Figure 3-4).

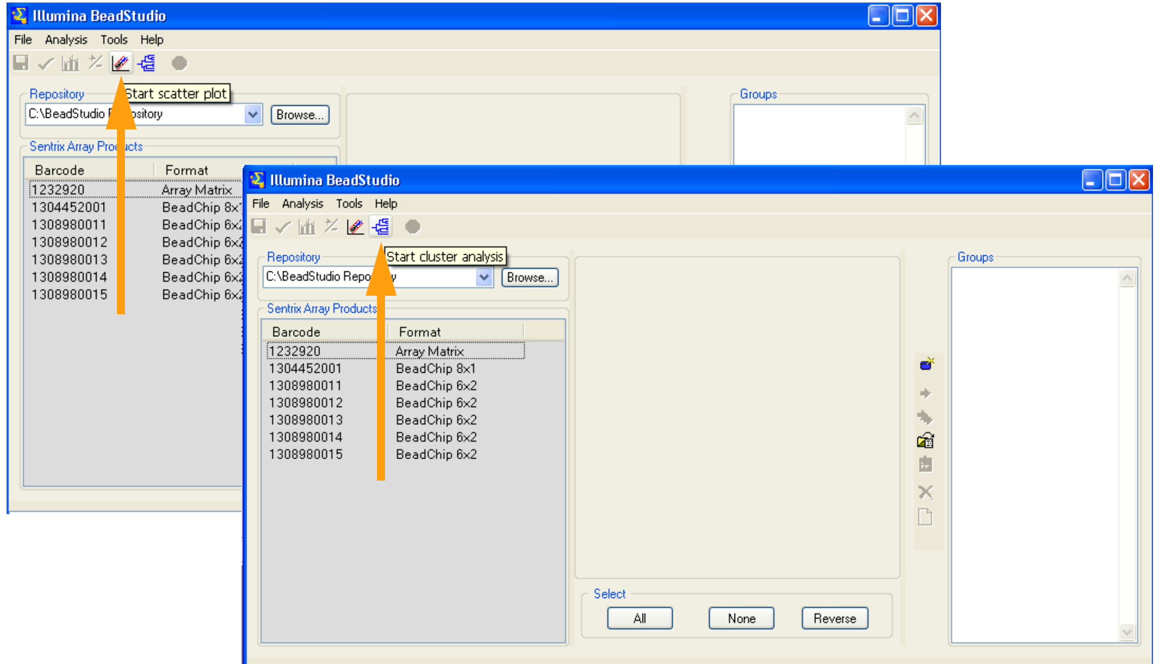


Figure 3-4 Start scatter plot and Start cluster analysis Tool Bar Buttons

b. Browse to your processed data output files, highlight the file you want to use (gene_profile, gene_probe_profile, or gene_diff), and click **Open**.

c. BeadStudio will launch the selected tool using the experimental data you specified in step b, above.

Scatter Plots

Once gene analysis or differential expression analysis has been completed, you can create Scatter Plots using BeadStudio output files.

NOTE:

You can generate scatter plots using either gene_profile, gene_probe_profile, or gene_diff files. The files must be column-oriented.

To create a Scatter plot:

1. Open the **Scatter plot** tool (see *Accessing the Data Visualization Tools* on page 3-3).
2. In the **Scatter Plot Source Data** dialog box (Figure 3-5), select the position you want for the XValues and click **>>>X Values.**
3. Next, select the position you want for the YValues, and click **>>>Y Values**.

NOTE:

For Differential Expression, the X value is pre-set and unchangeable.

4. Click **Create Scatter Plot** to view the plot.

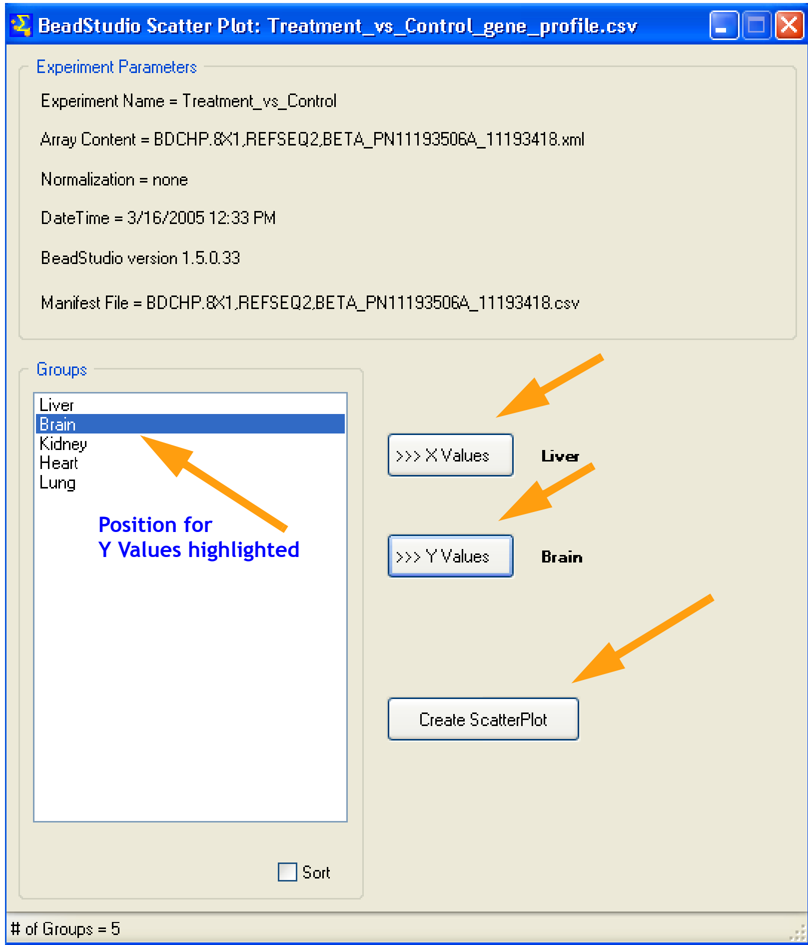


Figure 3-5 Scatter Plot Source Data Dialog Box

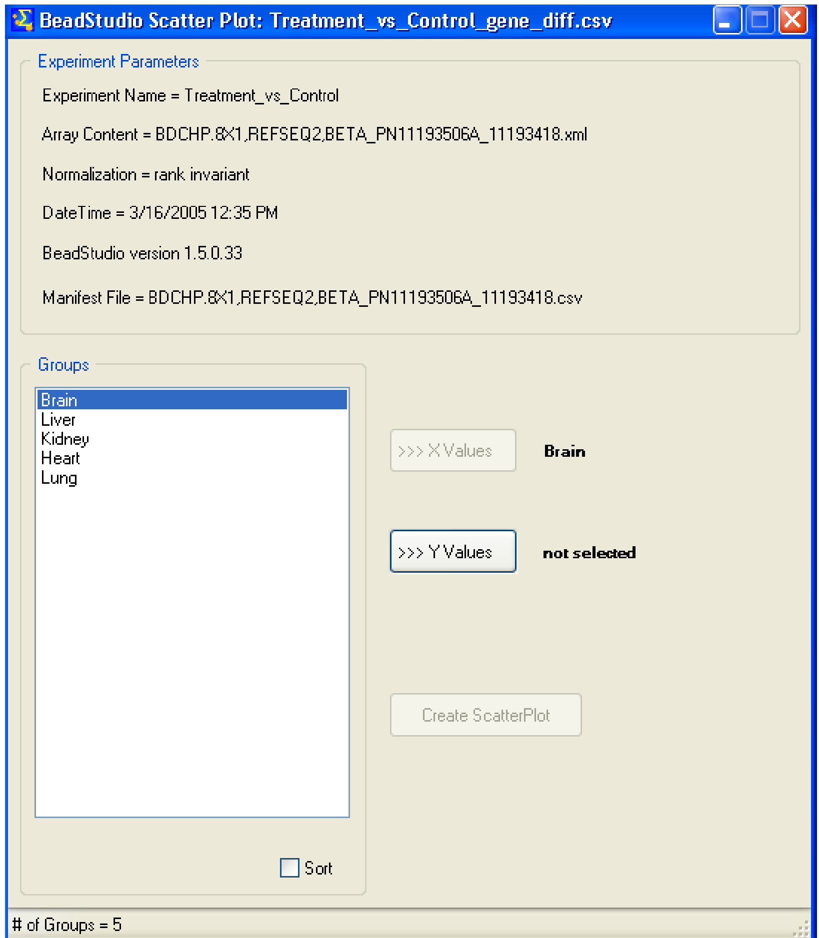


Figure 3-6 Source Data Dialog Box for Differential Analysis

NOTE:

The X Value is automatically assigned the RefGroup for Differential Analysis. The X Value cannot be changed.

5. The **Control Panel** (Figure 3-7) displays automatically (to hide or close the **Control Panel,** from the **Scatter Plot** menu bar, select **View | Control Panel**).

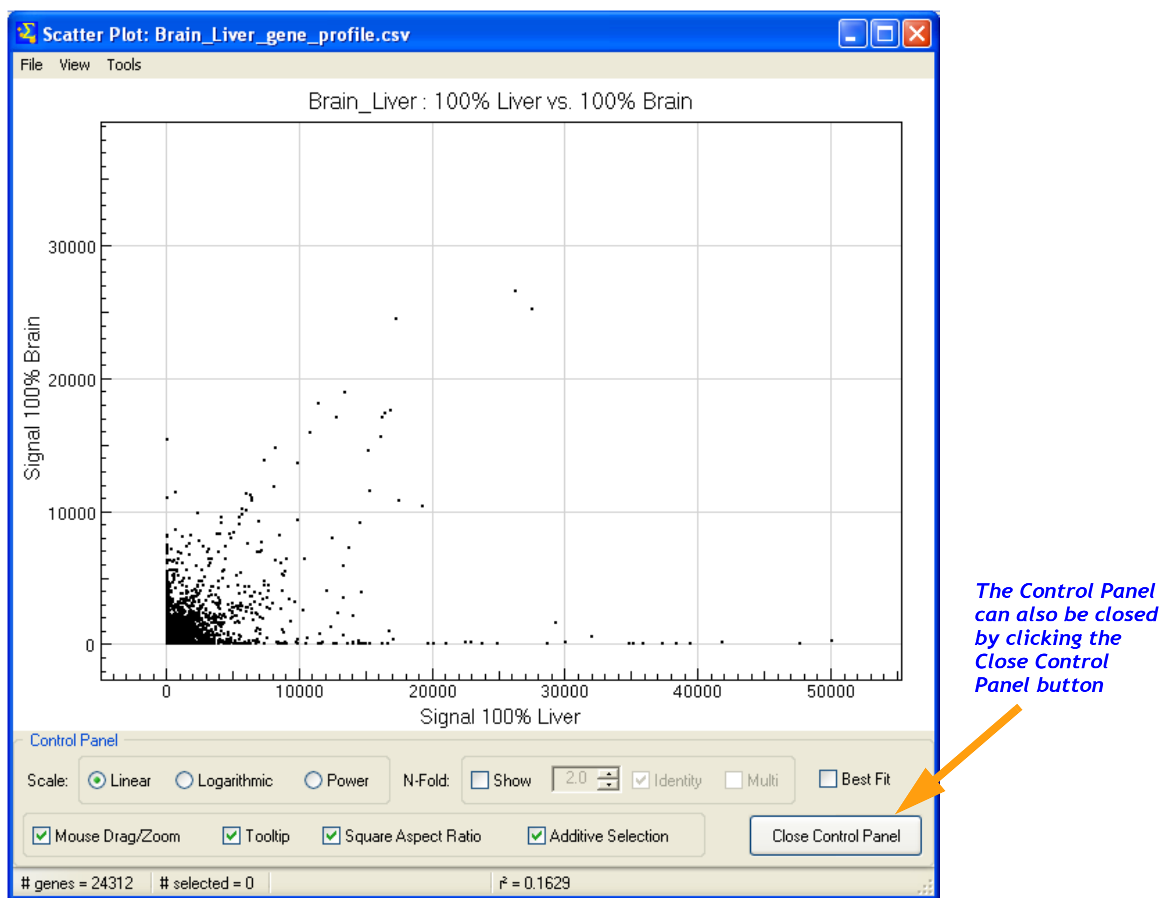


Figure 3-7 Scatter Plot

Table 3-1 lists and describes **Control Panel** functions.

Table 3-1 Scatter Plot Control Panel Descriptions

Item	Description
Scale	• **Linear** radio button *When enabled, X and Y axes are on a linear scale* • **Logarithmic** radio button *When enabled, X and Y axes are on a logarithmic scale* • **Power** radio button *When enabled, X and Y axes are on* n^{th} *root scale (where n is an odd number from 3 to 9)*
N-Fold	• **Show** checkbox *When checked, shows n-fold lines and allows you to select the fold value* • **N-fold setting selector** *When **Show** is checked, allows you to select the fold change* • **Identity** checkbox *When checked, BeadStudio displays the identity line in **bold red color**. If a gene is on this line, its X and Y intensities are equal.* • **Multi** checkbox *When checked, BeadStudio displays additional incremental fold change regions*

Table 3-1 Scatter Plot Control Panel Descriptions (continued)

Item	Description
Options	• **MouseDrag/Zoom** checkbox *When checked, allows you to drag and zoom in or out using the mouse. Use the mouse wheel to zoom in or out. If your mouse does not have a wheel:* ▹ *Press the Shift key while pressing the left mouse button.* ▹ *Drag to create a rectangle around an area to zoom in on.* ▹ *Release the Shift key and the mouse button to zoom.* ▹ *To return to normal view, from the Scatter Plot Tools menu (or the Scatter Plot context menu), select Auto Scale Axes.* • **Tooltip** checkbox *When checked, Scatter Plot displays gene symbol and X/Y intensities as mouse hovers over the gene.* • **Square Aspect Ratio** checkbox *When checked, X axis scale is equal to Y axis scale.* • **Additive Selection** checkbox *When checked, any new gene selection will be added to the scatter plot, along with previous selections. When not checked, any new selection replaces the previous selection(s).* • **Best Fit** checkbox *When checked, presents the Scatter Plot in the optimal fit for the genes of interest (linear equation is displayed in Control Panel next to* r^2 *values).* • **Close Control Panel** button *Click to close the Control Panel.*
Status Bar	• **# genes ( ) =** *Displays the number of genes visible in the Scatter Plot, and the total number of genes ( ).* • **# selected =** *Displays the number of selected genes in the Scatter Plot.* • **Position** pane *Displays current X/Y position of gene (mouse pointer) on the Scatter Plot.*
r^2	Square of the correlation coefficient. *NOTE: If the scatter plot is in linear scale, the* r^2 *value is calculated in linear space; if the scatter plot is in logarithmic scale,* r^2 *is calculated in log space.*

6. To use a variety of plot tools, from the menu bar, click **Tools** (Figure 3-8).

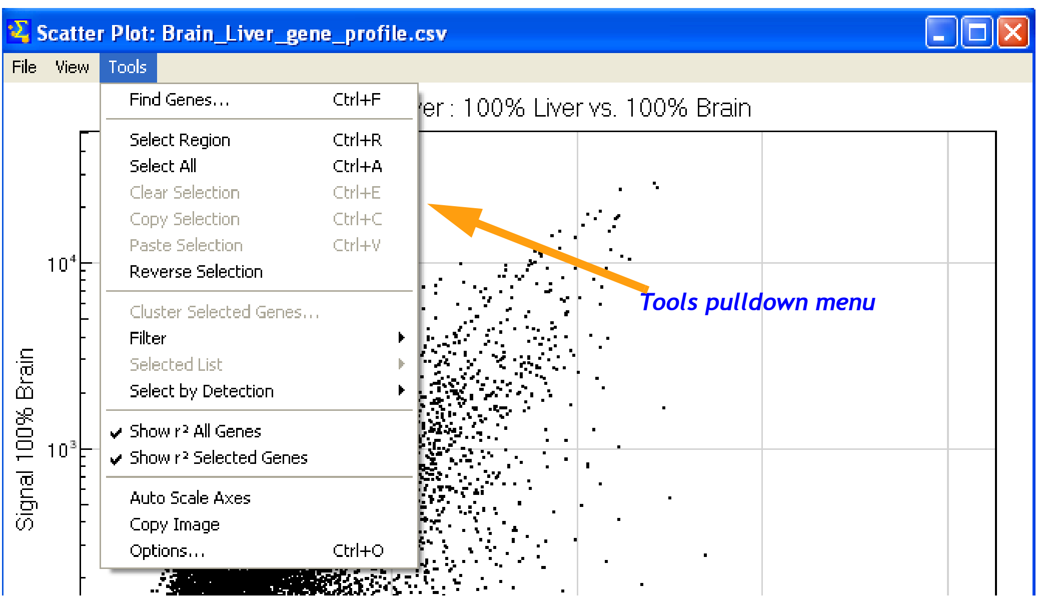


Figure 3-8 Scatter Plot Tools Menu

Table 3-2 describes the available tools.

Table 3-2 Scatter Plot Tools Menu Item Descriptions

Tool Name	Description
Find Genes...	When selected, opens the **Find Genes...** dialog box, where you can either: - Enter a list of genes separated by commas (see *Finding Genes in the Scatter Plot* on page 3-19). *OR* - Load a search gene list from a text file. *NOTE: Use this tool to search for other things, such as definitions, etc.*
Select Region	When selected, cursor becomes a crosshair you can use to draw a yellow boundary around any region in the Scatter Plot. All genes in this region become selected.
Select All	When selected, all genes in the Scatter Plot are selected. When selected, genes are displayed in the currently selected color.
Clear Selection	When selected, clears any previous selections.
Copy Selection	When selected, places any previous selections to the clipboard.
Paste Selection	When selected, pastes the contents of the clipboard to the current location. Genes can be copied/pasted across any scatter plots, dendrograms, and other applications (e.g., Notepad or Excel).
Reverse Selection	When selected, reverses the current selection (i.e., selected genes are unselected).
Cluster Selected Genes...	When selected, opens the **Cluster Analysis Experiment Parameters** window, allowing you to cluster based on data for selected genes.
Filter	When selected, allows you to filter genes shown on the scatter plot by first selecting genes of interest, then clicking either **Selected Genes** or **Unselected Genes**. You can repeat filtering as many times as desired.
Selected List	When selected, allows you to: - Save the selected gene list as a text file - View the selected gene in a Web browser - Show Gene Symbols
Select by Detection	When selected, opens the BeadStudio **Select Detected Gene** dialog box, where you can select your desired detection level. Enabled only when detection data is available.
Select by Diff Score	When selected, allows you to select genes by their diff scores, using data from the gene_diff output file.

Table 3-2 Scatter Plot Tools Menu Item Descriptions (continued)

Tool Name	Description
Show r^2 All Genes	When checked, each time the number of all genes changes, BeadStudio will automatically recalculate the r^2. When unchecked, no automatic recalculation occurs and no r^2 is displayed.
Show r^2 Selected Genes	When checked, each time the number of selected genes changes, BeadStudio will automatically recalculate the r^2. When unchecked, no automatic recalculation occurs and no r^2 is displayed.
Auto Scale Axes	When selected, automatically scales the Scatter Plot X and Y axes.
Copy Image	When selected, copies the current image to the clipboard.
Options...	Opens the **ScatterPlot Options** dialog box, in which you can set: - Axes min/max values ▹ *Displays the minimum and maximum X and Y axis values. When Square Aspect Ratio is NOT checked, you can set new X and Y axis values.* - Text and fonts for Scatter Plot labels - Data Point size and scale - Power for the scale (odd number, from 3 to 9, with 3 as the default) - Colors for the Scatter Plot: ▹ *Axes* ▹ *Background* ▹ *Grid* ▹ *Data Points* ▹ *Selection*

Scatter Plot Context Menu Functions

Right-click anywhere in the Scatter Plot to view the context menu (Figure 3-9). The context menu contains features that can be applied to the selected object.

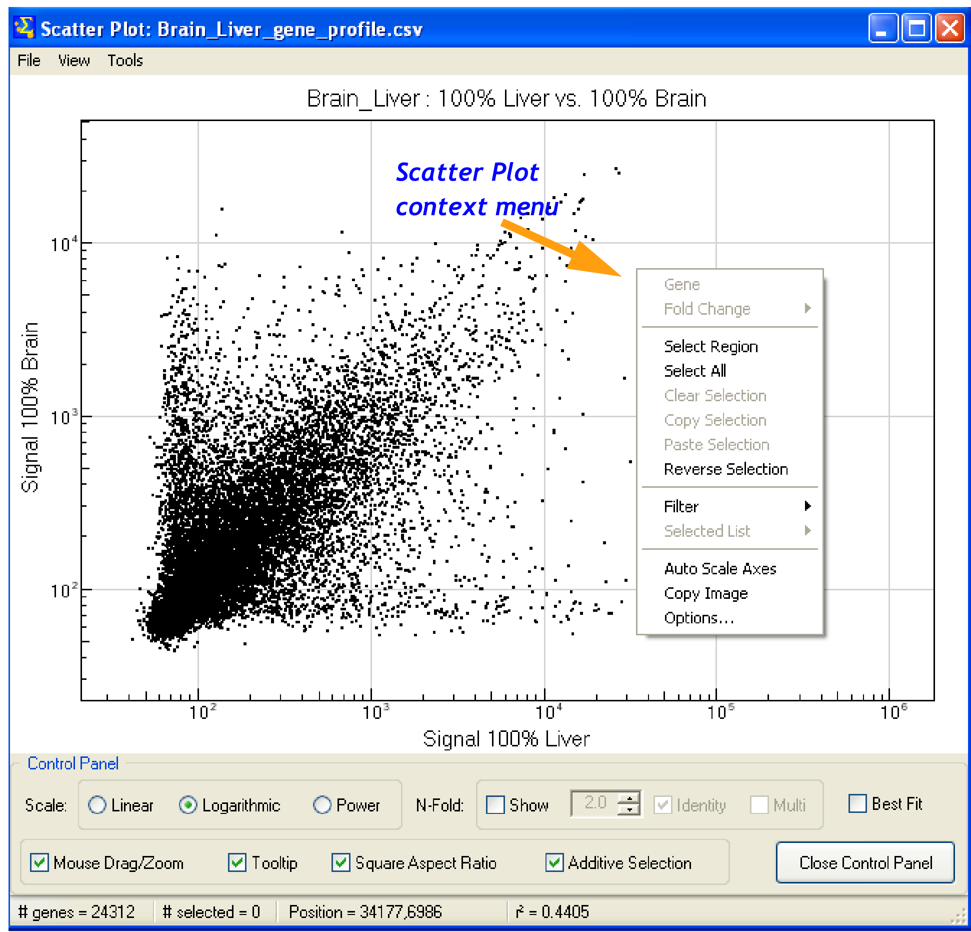


Figure 3-9 Scatter Plot Context Menu

Table 3-2 lists context menu items and their functions.

Table 3-3 Scatter Plot Context Menu Item Descriptions

Item	Description
Gene	When a gene is selected, displays the gene symbol.
Fold Change	If fold change lines are present, displays the fold change limits for current cursor location. Allows you to select/de-select all genes inside the fold change.
Select Region	When selected, cursor becomes a crosshair you can use to draw a yellow boundary around any region in the Scatter Plot.
Select All	When selected, all genes in the Scatter Plot are selected and displayed in the currently selected color.
Clear Selection	When selected, clears the existing selection.
Copy Selection	When selected, places the existing selection on the clipboard.
Paste Selection	When selected, pastes the contents of the clipboard to the current location.
Reverse Selection	When selected, reverses the current selection (i.e., selected genes are unselected).
Filter	When selected, allows you to filter genes shown on the scatter plot by first selecting genes of interest, then removing unselected genes from the scatter plot. You can repeat these steps as many times as desired.
Selected List	When selected, allows you to: - Save the selected gene list as a text file - View the selected gene in a Web browser - Show Gene Symbols

Table 3-3 Scatter Plot Context Menu Item Descriptions (continued)

<table>
<tr><th>Item</th><th>Description</th></tr>
<tr><td>Auto Scale Axes</td><td>When selected, automatically scales the Scatter Plot X and Y axes.</td></tr>
<tr><td>Copy Image</td><td>When selected, places the Scatter Plot image on the clipboard.</td></tr>
<tr><td>Options...</td><td>Axes
Displays the minimum and maximum X and Y axis values. When Square Aspect Ratio is NOT checked, you can set new X and Y axis values.
Labels
Allows you to choose font properties for the Scatter Plot title, X axis, and Y axis.
Data Points
Allows you to select a point size and style for the Scatter Plot data points.
Scale
Allows you to select a power (3, 5, 7, or 9) for the Power setting.
Colors
Click in each box to bring up the color palette and set colors for:
Axes
Background
Grid
Data Points
Selection</td></tr>
</table>

Finding Genes in the Scatter Plot

BeadStudio provides a path to gene property information, including gene ID, intensities, and gene ontology information.

1. From the Scatter Plot menu bar, select **Tools | Find Genes...** (Figure 3-10).

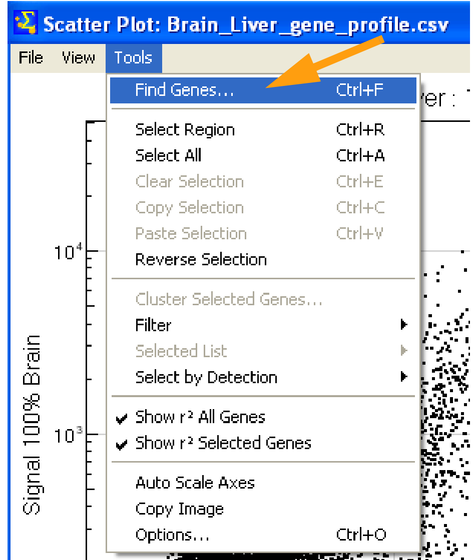


Figure 3-10 Find Genes... Tool Selected

2. At the **Find Genes** dialog box (Figure 3-11), select specific genes based on the following fields in their GenBank database records:
   - *Gene ID*
   - *Symbol*
   - *Synonyms*
   - *Accession*
   - *Definition*
   - *Ontology*

   In the **Search in** pane, select the radio button next to the GenBank database field you wish to search.

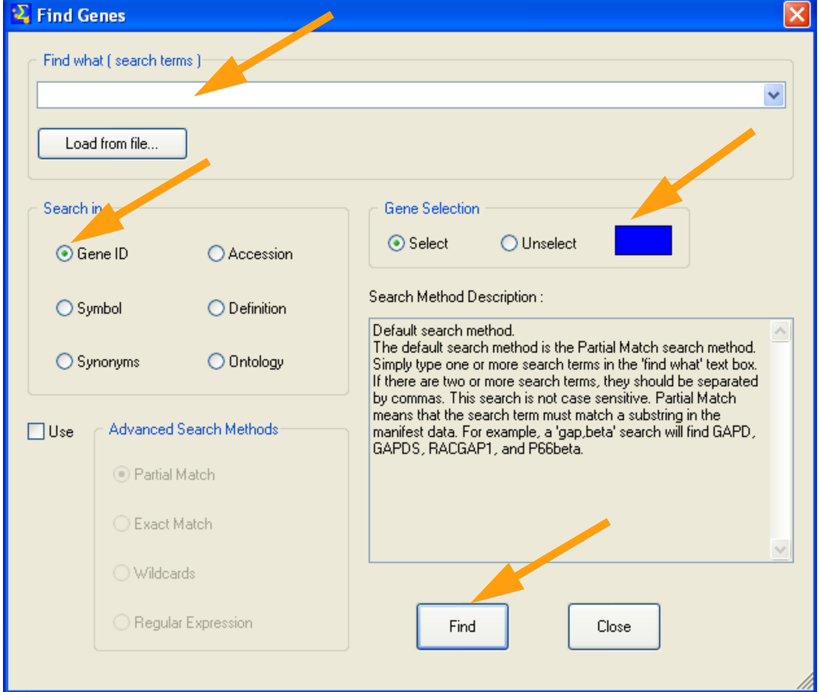


Figure 3-11 Find Genes Dialog Box

3. In the **Find what** text field, enter the search text.

NOTE:

By default, searches are partial. For example, if you search the word 'VEGF' in the Symbol field, the search will return not only VEGF, but also, VEGFB and VEGFC.

Multiple search terms can be used, separated by commas.

Search terms can also be loaded from a text file. The file should have each term on a separate line.

4. Either:
   - *Click **Select** to select found genes using the current selection color (default).*
     - *You can change selection color by clicking the color box in the **Gene Selection** pane.*
   - *Click **Unselect** to unselect found genes that were previously selected*
5. Click **Find** to return to the Scatter Plot with the identified genes identified highlighted.
6. For more advanced search options, check the **Use** box next to the **Advanced Search Methods** pane. These methods are described in the '**Search Method Description**' box (Figure 3-11).

7. The Scatter Plot will display the selected gene in the color you selected (Figure 3-12). Zoom in for a better view.

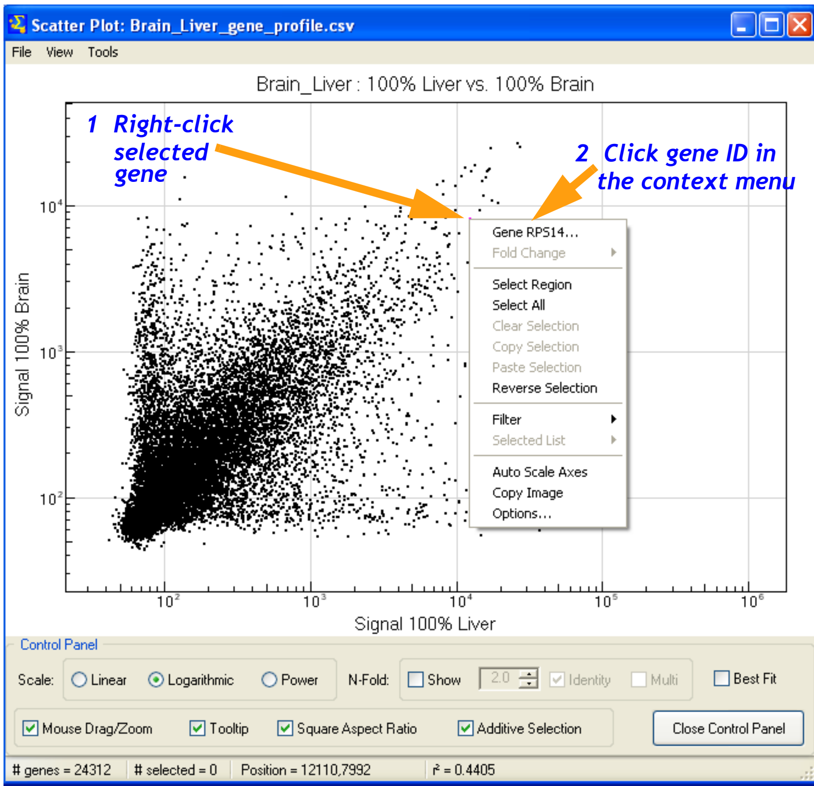


Figure 3-12 Zoom in to See Selected Genes

8. To bring up the **Gene Properties:** dialog:
   a. Right-click the selected gene.
   b. Click **Gene Symbol** in the context menu.

The following paragraphs illustrate the **Gene Properties:** dialog box functions.

Data Tab

Figure 3-13 illustrates the **Gene Properties:** window **Data** tab.

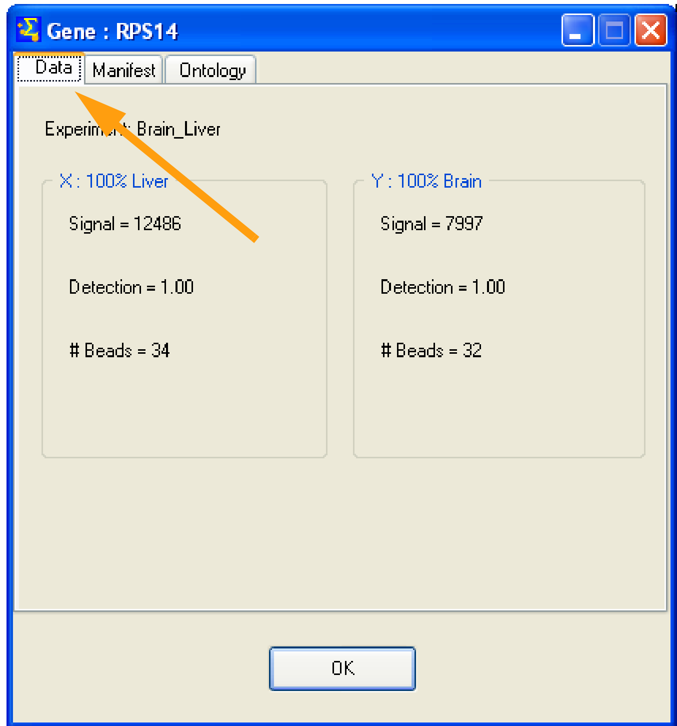


Figure 3-13 Gene Properties: Window Data Tab

Figure 3-14 illustrates the **Gene Properties:** window Data tab for differential expression.

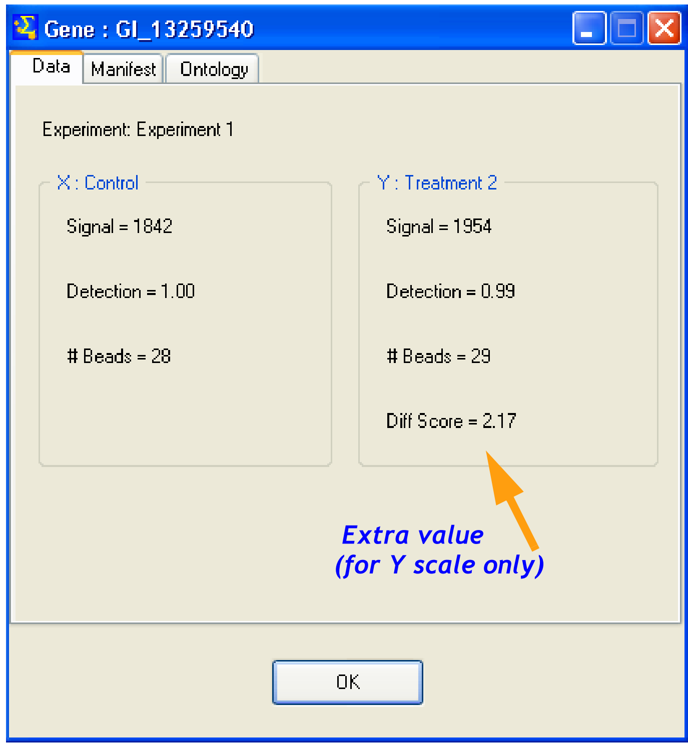


Figure 3-14 Gene Properties: Window Data Tab

Manifest Tab

Figures 3-15, 3-16, and 3-17 illustrate functions of the **Manifest** tab.

1. When you click the Accession link (Figure 3-15), BeadStudio jumps to the National Center for Biotechnology Information (NCBI) website (Figure 3-16) where you can view the record for the selected gene.

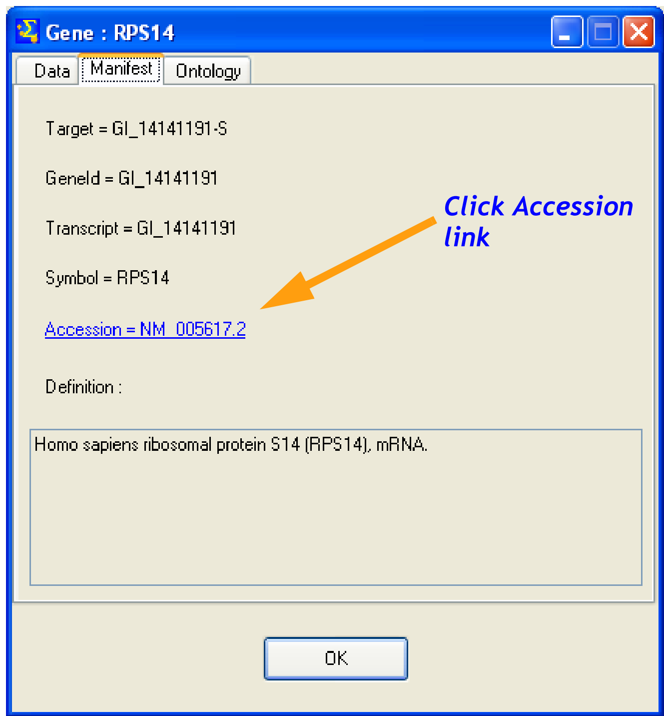


Figure 3-15 Gene Properties: Window Manifest Tab

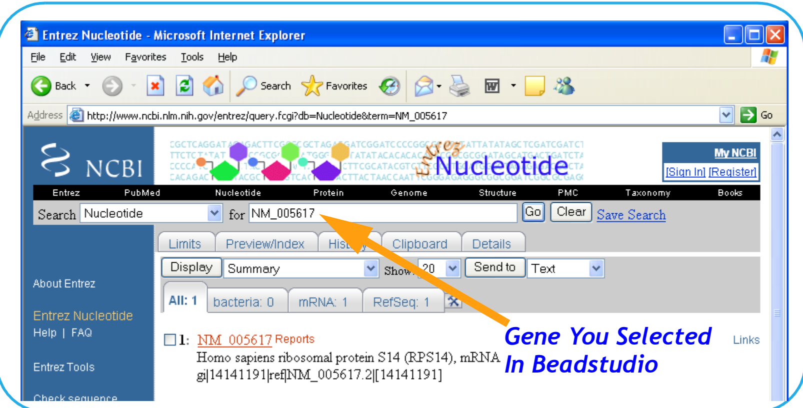


Figure 3-16 NCBI Website

2. Click the gene name to see the GenBank record (Figure 3-17).

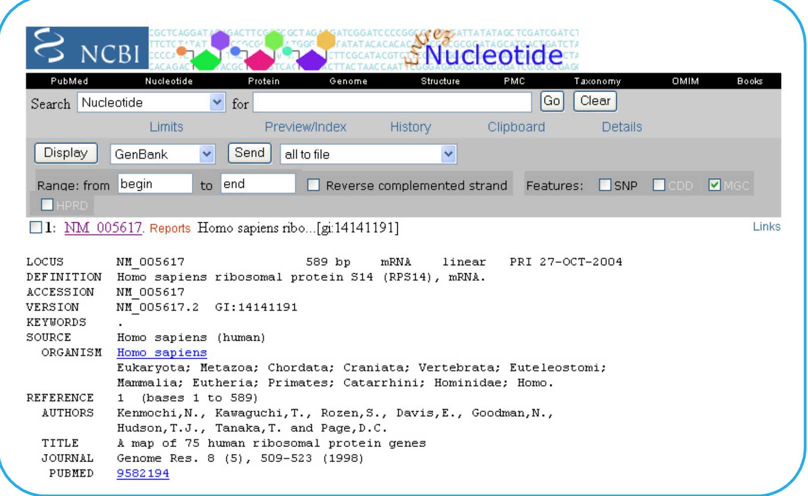


Figure 3-17 NCBI Record

Ontology Tab

Figure 3-18 illustrates the **Ontology** tab. This tab provides a quick reference to the NCBI gene ontology information.

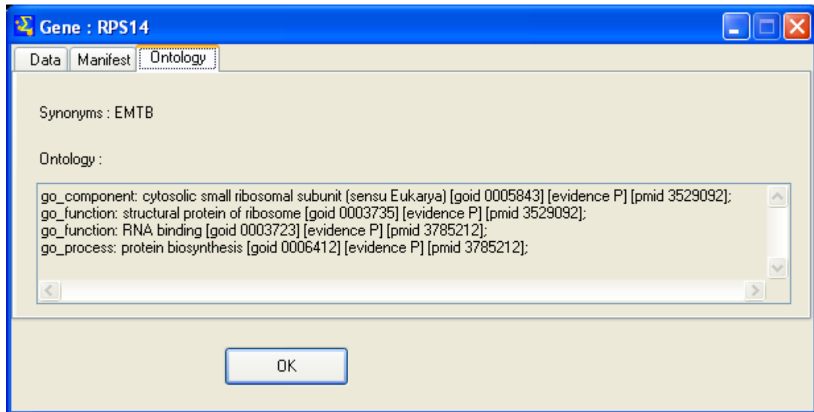


Figure 3-18 Gene Properties: Ontology Tab

Other Scatter Plot Functionalities

- Click and drag to move the Scatter Plot around.
- Shift-click to zoom in to a particular region.
- Rotate mouse wheel to zoom in and out.
- Control-click, hold, and drag to select a specific gene or group of genes.

Cluster Analysis

Clustering is an analysis method used to group sets of objects into subsets or clusters. Objects assigned to the same cluster are more closely related to one another than objects assigned to different clusters. In the context of gene expression, the method can be used to answer two basic questions:

- Which genes show similar patterns of gene expression across a series of samples?

  Useful for identifying genes in common pathways, or genes that coordinately respond to the same stimuli.

- Which samples are most similar based on the expression levels of genes within them?

  Useful for identifying conditions that generate a common metabolic response. For example, in a toxicology study, if an unknown compound induces a pattern of expression similar to that induced by a panel of genotoxins, it is likely that the unknown is a genotoxin.

Mathematicians have devised dozens of clustering methods using different metrics to establish relationships between sets of values. In BeadStudio, clustering occurs using the agglomerative nesting with average linkage method. BeadStudio offers four clustering metrics for calculating dissimilarities:

- Correlation

  Computes the Pearson correlation using a 1 - r distance measure

- Absolute Correlation

  Computes the Pearson correlation using a 1 - |r| distance measure

- Manhattan

  Computes the distance between two points if a grid-like path is followed

- Euclidean

  Computes the shortest distance between two points

NOTE:

Generally, Illumina recommends using multiple clustering methods to validate results. Groupings with a true biological basis will usually show up regardless of the algorithm used.

Similarities and Distances

The first decision to be made is how “similarity” is to be defined. There are several ways to compute the similarity of two series of numbers. The most commonly used similarity metric is the Pearson correlation. The Pearson correlation coefficient between any two series of numbers $X = \{X_1, X_2, \ldots X_N\}$ and $Y = \{Y_1, Y_2, \ldots Y_N\}$

is defined as:

$$r = \frac{1}{N}\sum_{i=1,N}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right)$$

Distance is then defined as $1 - r$ for Correlation and $1 - |r|$ for Absolute Correlation. BeadStudio also uses Manhattan $(\Sigma|X_l - Y_l|)$ and squared Euclidean $\left(\Sigma(X_l - Y_l)^2\right)$ distances.

BeadStudio presents the clustering information in the form of a dendrogram, a tree-like structure whose branches correspond to genes or samples, depending on how the analysis is run. The distance on the X axis establishes the similarity relationships among the genes or samples. For example, if the dendrogram plots the similarity of samples based on gene expression, samples C and D are very similar to each other, less similar to B, and even less similar to A (Figure 3-19).

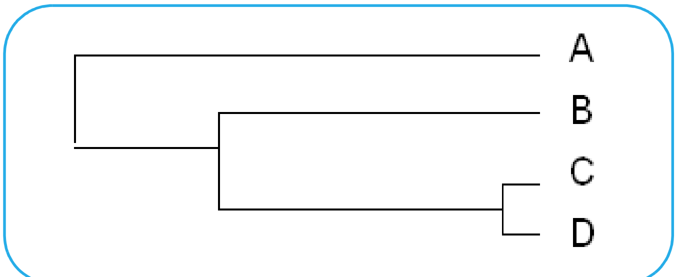


Figure 3-19 Dendrogram, Similarity Example

After clustering, nodes are reordered starting near the top to ensure that node “ar” is closer to “B” than node “al”, and node “bl” is closer to “A” than node “br” (Figure 3-20).

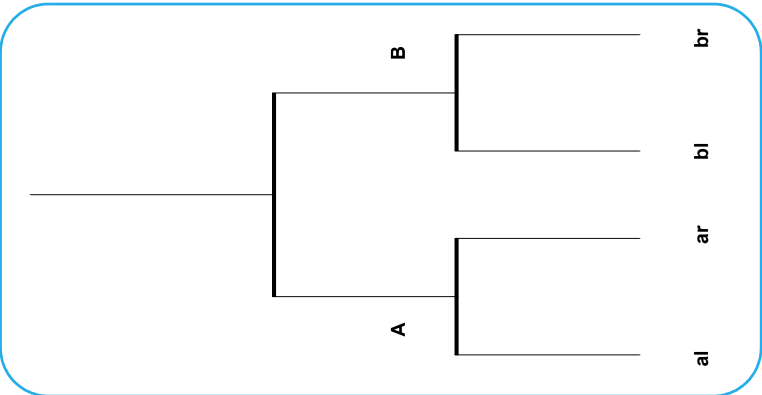


Figure 3-20 Dendrogram, Showing Nodes

Analyze Clusters

To analyze clusters:

1. Open the **cluster analysis** tool (see *Accessing the Data Visualization Tools* on page 3-3).
2. In the **Cluster Analysis:** dialog box (Figure 3-21):
   a. **Groups** pane -- highlight the group(s) whose clusters you wish to analyze (or click **Select All**).
   b. Enable the **Sort** checkbox to sort the items in “groups” list box alphabetically in ascending order.
   c. **Cluster** pane -- Click **Genes** or **Samples**. If **Genes** is selected, the dendrogram will display a cluster of genes. If **Samples** is selected the dendrogram will display a cluster of samples.

NOTE:

Clustering **Samples** is much faster than clustering **Genes**. Clustering many genes (thousands) can take hours.

d. **Metric** pane -- Select the metric you would like BeadStudio to use to calculate clusters.

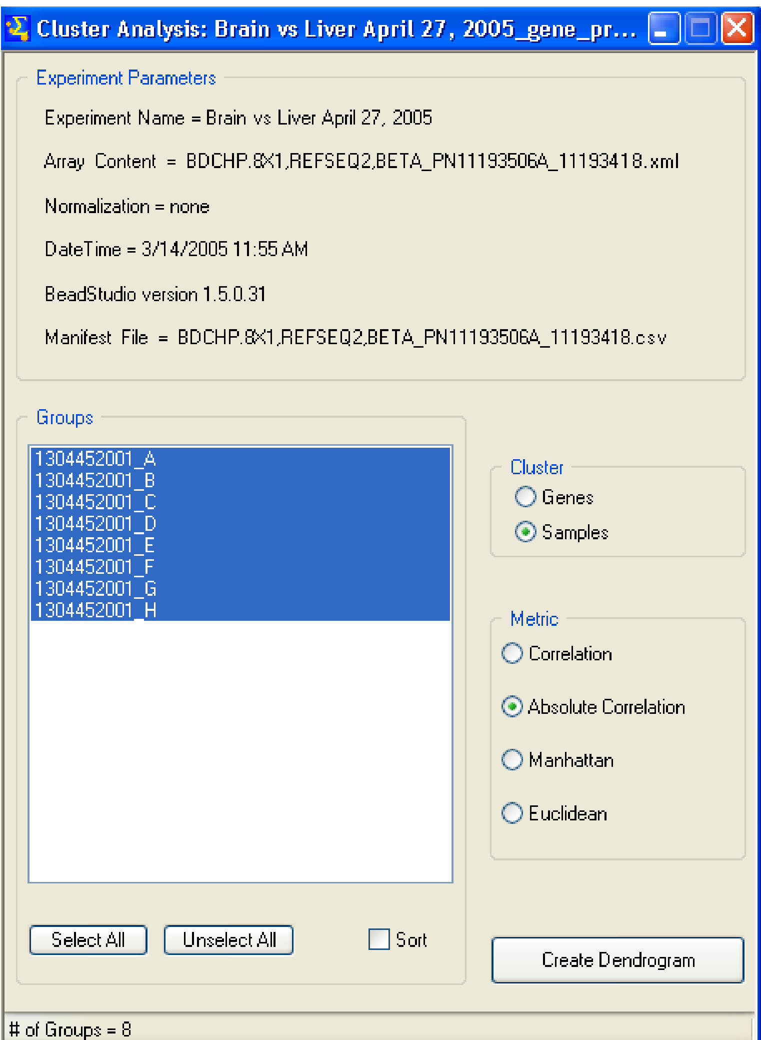


Figure 3-21 Cluster Analysis Dialog Box

e. Click **Create Dendrogram** to view the graph (Figure 3-22).

NOTE:

The scale at the bottom of the dendrogram shows dissimilarity between modes. See *Similarities and Distances* on page 3-29.

Figure 3-22 Dendrogram

3. Right click in the dendrogram to view the context menu (Figure 3-23).

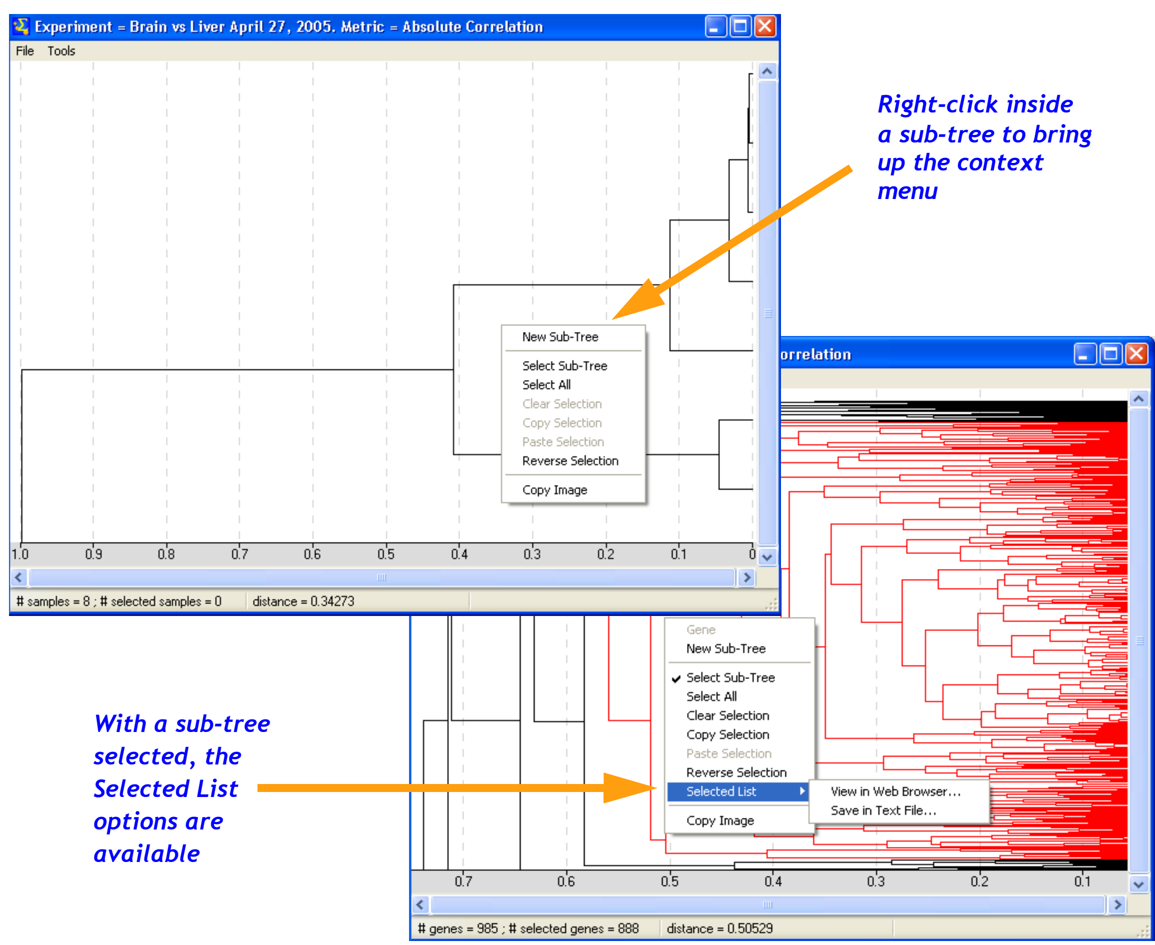


Figure 3-23 Dendrogram with Context Menu

Dendrogram Context Menu Selections

Table 3-4 lists and describes the Dendrogram context menu items.

Table 3-4 Dendrogram Context Menu Descriptions

Item	Description
New Sub-Tree	When selected, displays the selected sub tree in a new window. This feature is disabled when the cursor is outside of any tree.
Select Sub-Tree	When selected, highlights the sub-tree in green. This feature is disabled when the cursor is outside of any tree.
Select All	Click to select all sub-trees.
Clear Selection	Clears any selection.
Copy Selection	Copies current selection(s) to the clipboard.
Paste Selection	Pastes current clipboard contents to the location you choose.
Reverse Selection	Reverses the last selection made.
Selected List	When a sub-tree is selected, click to select one of these two options: - **View in Web Browser...** - **Save in Text File...**
Copy Image	Copies current image to the clipboard.

View the Sub-Tree List Directly in the Dendrogram

To view the sub-tree list directly in the dendrogram, zoom in. The sub-tree list appears to the right of the dendrogram (Figure 3-24).

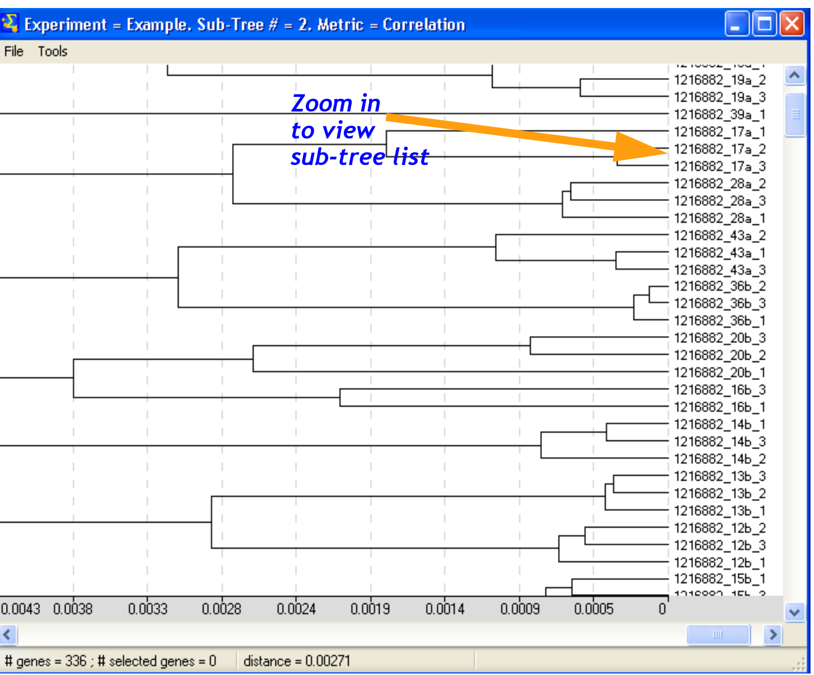


Figure 3-24 Zoom in to View Sub-Tree List

Copy/Paste Clusters

You can copy/paste gene clusters from a Scatter Plot to a dendrogram and vice versa. Refer to Figures 3-25 thru 3-28.

From Scatter Plot to Dendrogram

1. To select clusters for copying from the Scatter Plot,
   - *Select **Tools** | **Select Region** from the pulldown menu.*

     OR
   - *Open the context menu and click **Select Region**.*
2. Using the crosshair tool, draw around the genes you wish to copy.

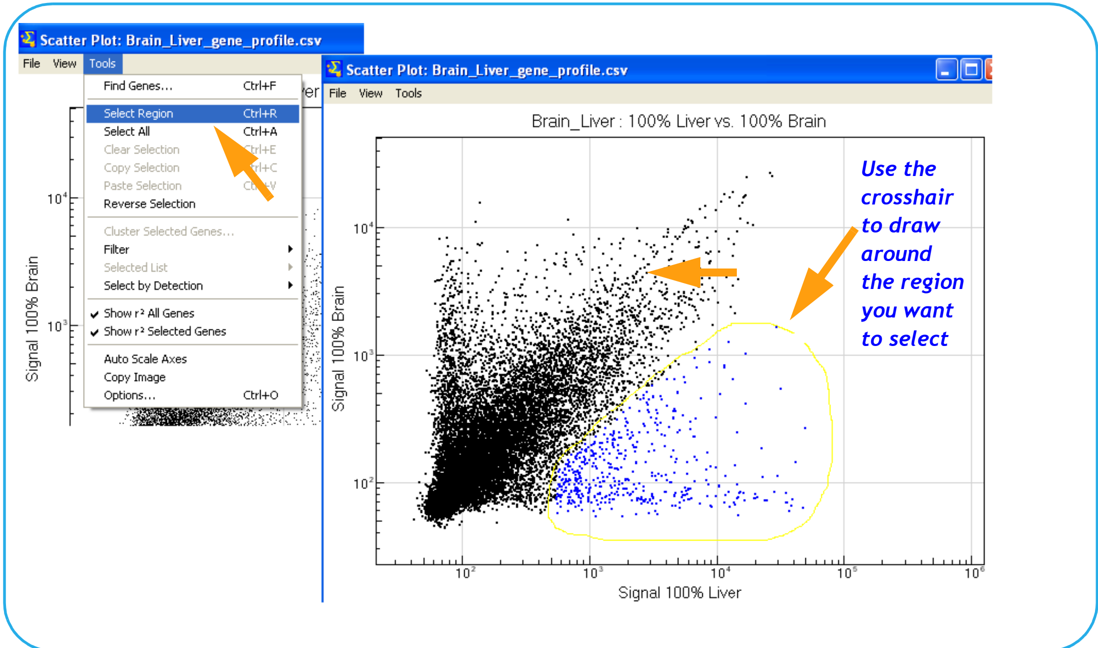


Figure 3-25 Select Region

NOTE:

The selected genes will change color (blue by default, or the color you have set in **Scatter Plot Options**).

3. To copy the selection to the clipboard:
   - *Select **Tools | Copy Selection** from the pulldown menu.*

     OR
   - *Open the context menu and click **Copy Selection**.*
4. To paste the selection into the dendrogram:
   - *Select **Tools | Paste Selection** from the pulldown menu.*

     OR
   - *Open the context menu and click **Paste Selection**.*

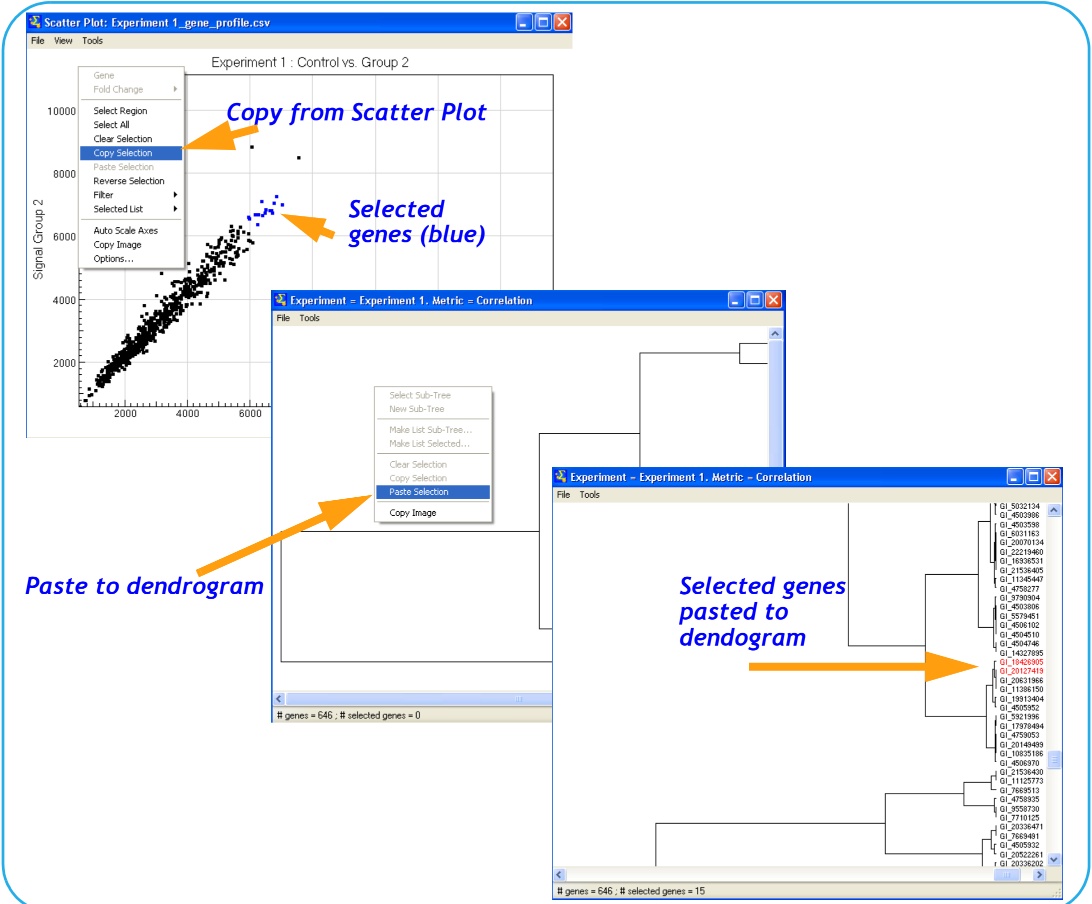


Figure 3-26 Copy & Paste

From Dendrogram to Scatter Plot

1. To select clusters for copying from the dendrogram, open the context menu and click **Select Sub-Tree**.

NOTE:

Be sure to click INSIDE the sub-tree you want to select. The selected sub-tree will turn red.

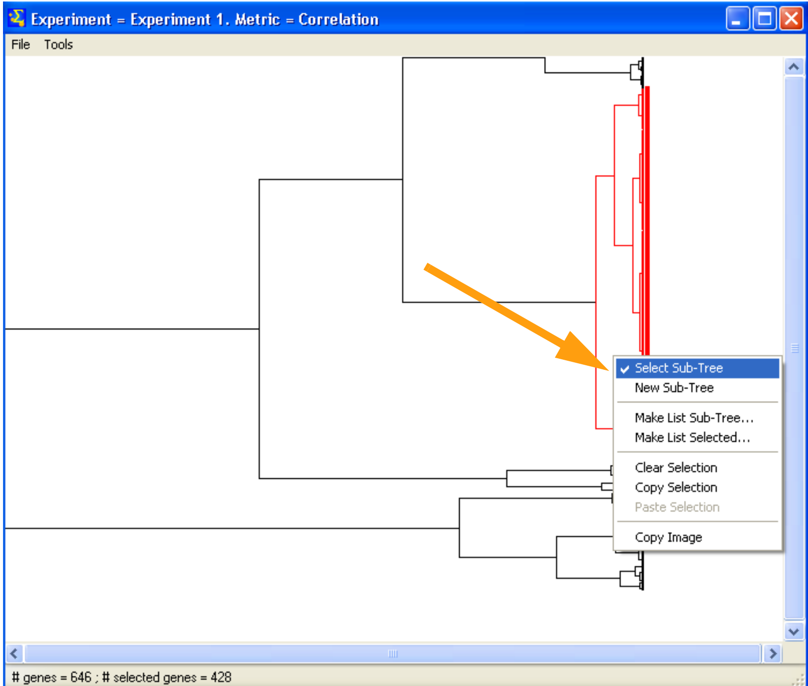


Figure 3-27 Select Sub-Tree

2. From the context menu, click **Copy Selection**.
3. To paste selection into the Scatter Plot:
   - *Select **Tools | Paste Selection** from the pulldown menu.*

     OR
   - *Open the context menu and click **Paste Selection**.*

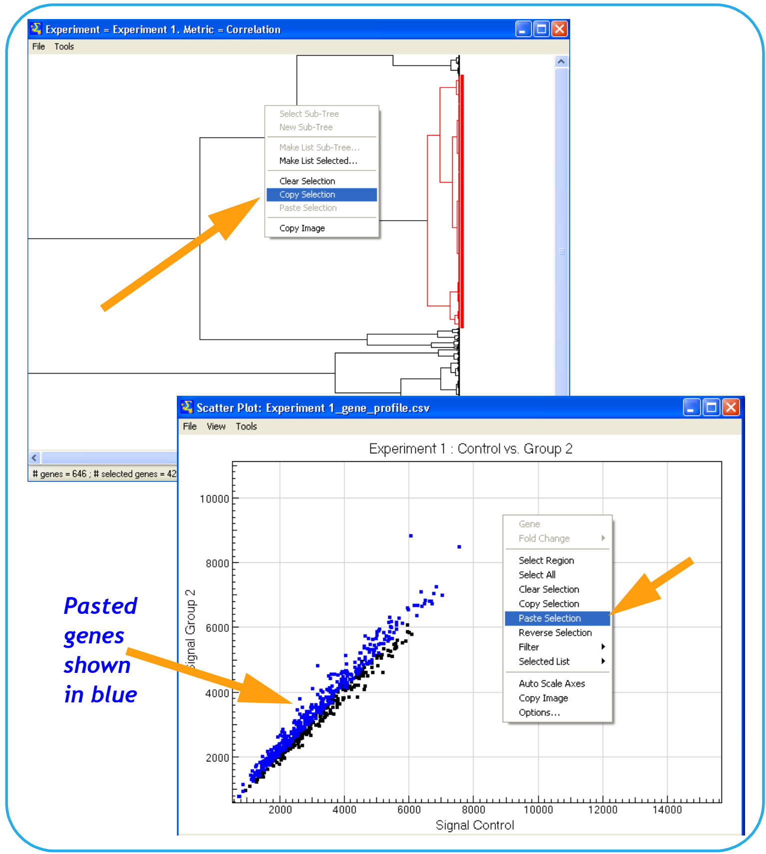


Figure 3-28 Copy & Paste

Control Summary Reports

The two following sections describe Control Summary Reports for DirectyHyb and the DASL™ Assay.

Control Summary Reports for DirectHyb

BeadStudio can displays a graphic Control Summary Report for the selected samples based on performance of the built-in controls (Figure 3-29).

For more detailed information on the controls, see the *System Controls* appendix in the appropriate Illumina product manual.

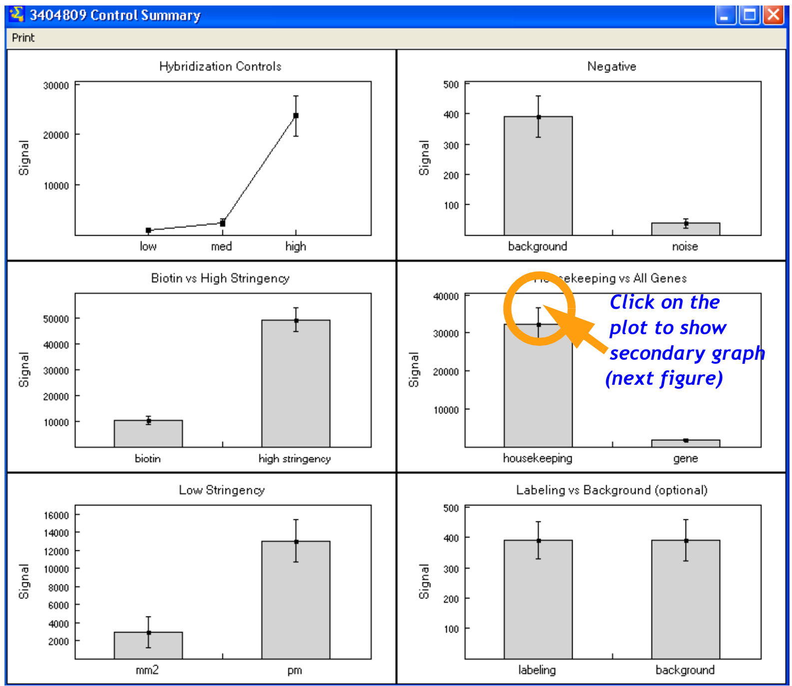


Figure 3-29 Control Summary Report

4. To view secondary graph(s), click on a data point in any of the graphs shown in Figures 3-29. Figure 3-30 illustrates an example of the secondary graph for housekeeping controls. Each point in the secondary graph represents a sample.

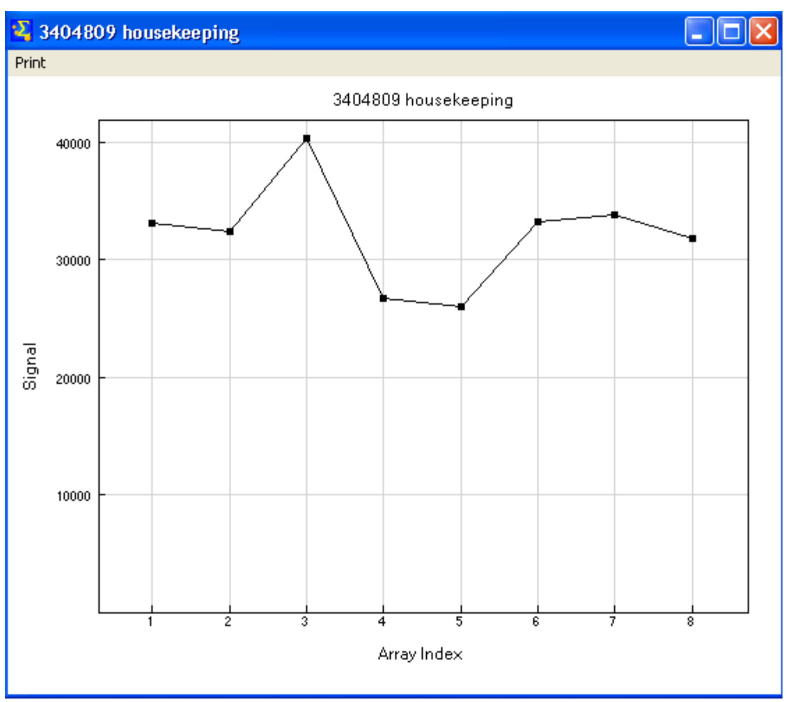


Figure 3-30 Housekeeping Controls Secondary Graph

5. To copy, change the page setup, or see a print preview, right click in any graph to use the context menu (Figure 3-31). To print the graph, either right click and use the context menu, or click **Print** in the upper right-hand corner of the menu bar.

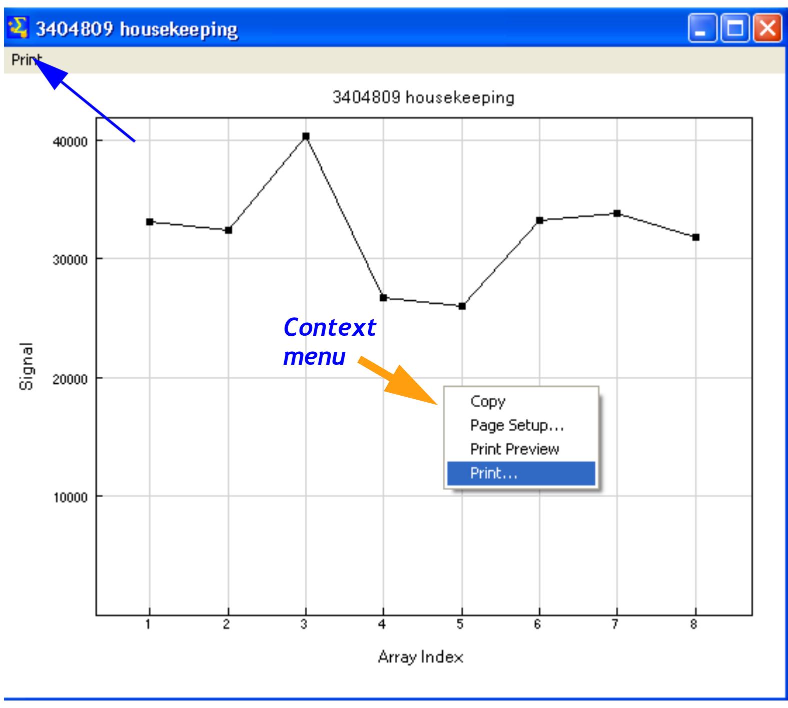


Figure 3-31 Control Summary Context Menu

Control Summary Reports for the DASL™ Assay

BeadStudio can display a graphic Control Summary for the selected samples based on performance of the built-in controls (Figure 3-32).

For more detailed information on the controls, see the "System Controls" appendix in the appropriate Illumina product manual.

Figure 3-32 Control Summary Reports

6. To view secondary graph(s), click on a data point in any of the graphs shown in Figure 3-32. Figure 3-33 illustrates an example of the secondary graph for contamination controls. Each point in the secondary graph represents a sample.

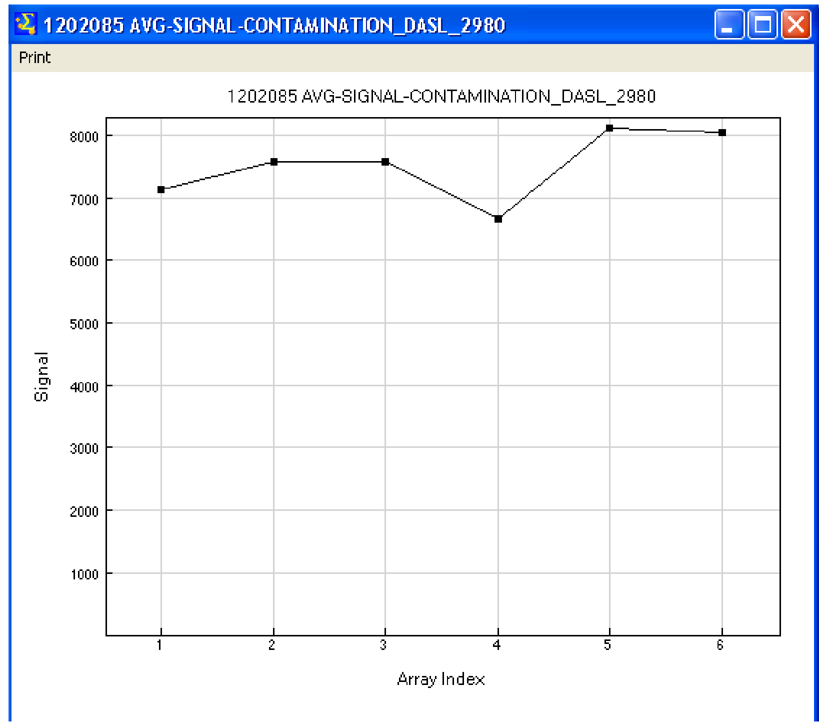


Figure 3-33 Contamination Controls Secondary Graph

7. To copy, change the page setup, or see a print preview, right click in any graph to use the context menu (Figure 3-34). To print the graph, either right click and use the context menu, or click **Print** in the upper right-hand corner of the menu bar.

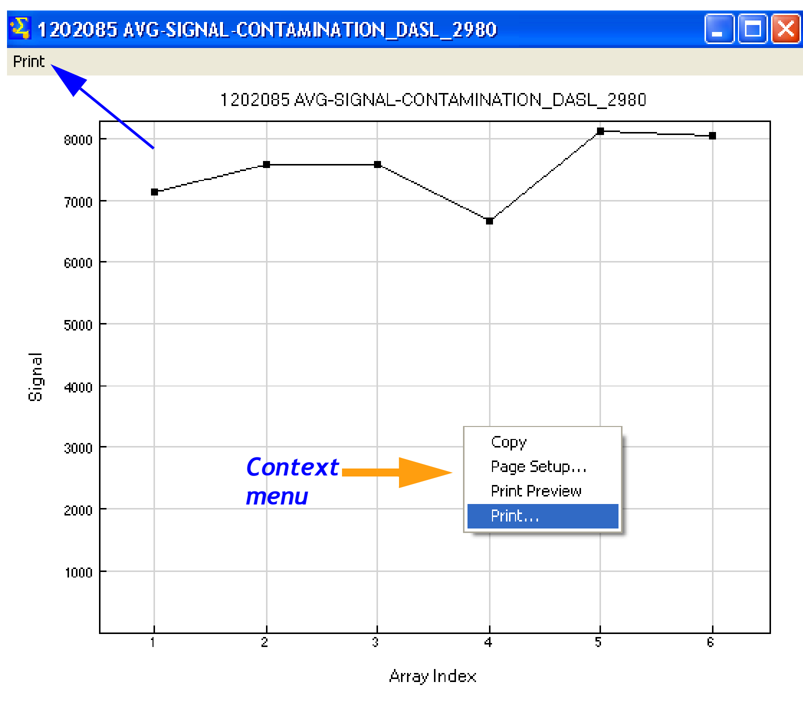


Figure 3-34 Control Summary Context Menu

Viewing Control Reports

To view control reports for any experiment at any time after analysis, open the **Control Summary Report** tool (see *Accessing the Data Visualization Tools* on page 3-3).

NOTES

Image Viewing

Topics

- Introduction *4-2*
- Using the Image Viewer *4-3*

Introduction

You can visually inspect any sample using the BeadStudio Image Viewer. Image Viewer allows you to view images and determine whether or not you want to include a particular sample(s) in your experiment.

In Image Viewer, you can:

- *view registration information for individual samples*
- *adjust contrast*
- *see registration for BeadStudio-processed images*
- *zoom in or out*
- *see pixel intensities*

Using the Image Viewer

Right-click on any sample to activate Image Viewer and view the BeadChip or Sentrix Array Matrix (SAM) image (Figures 4-1 and 4-2).

Table 4-1 lists and describes the Image Viewer features.

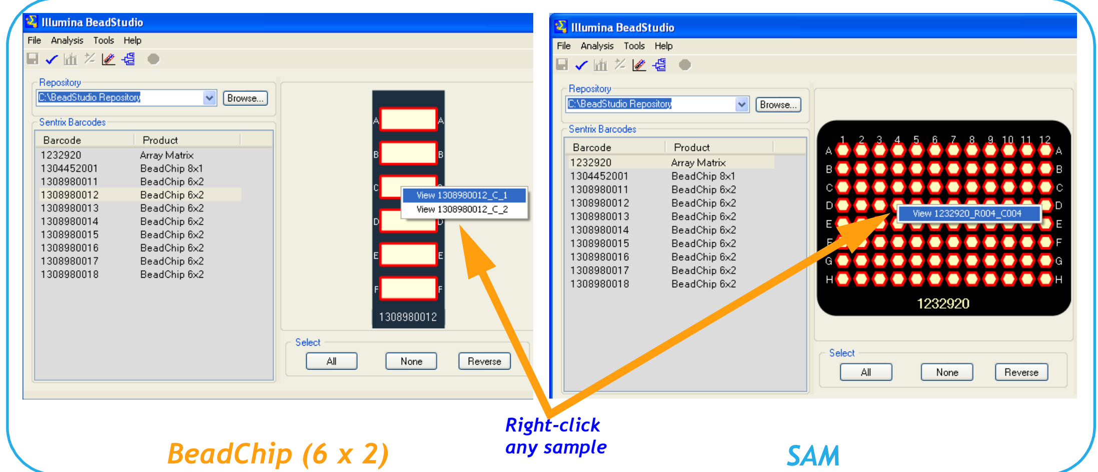


Figure 4-1 Right-Click any Sample to Activate Image Viewer

NOTE:

6 x 2 BeadChip images contain two images per sample section. After right-clicking, you must select which image to view.

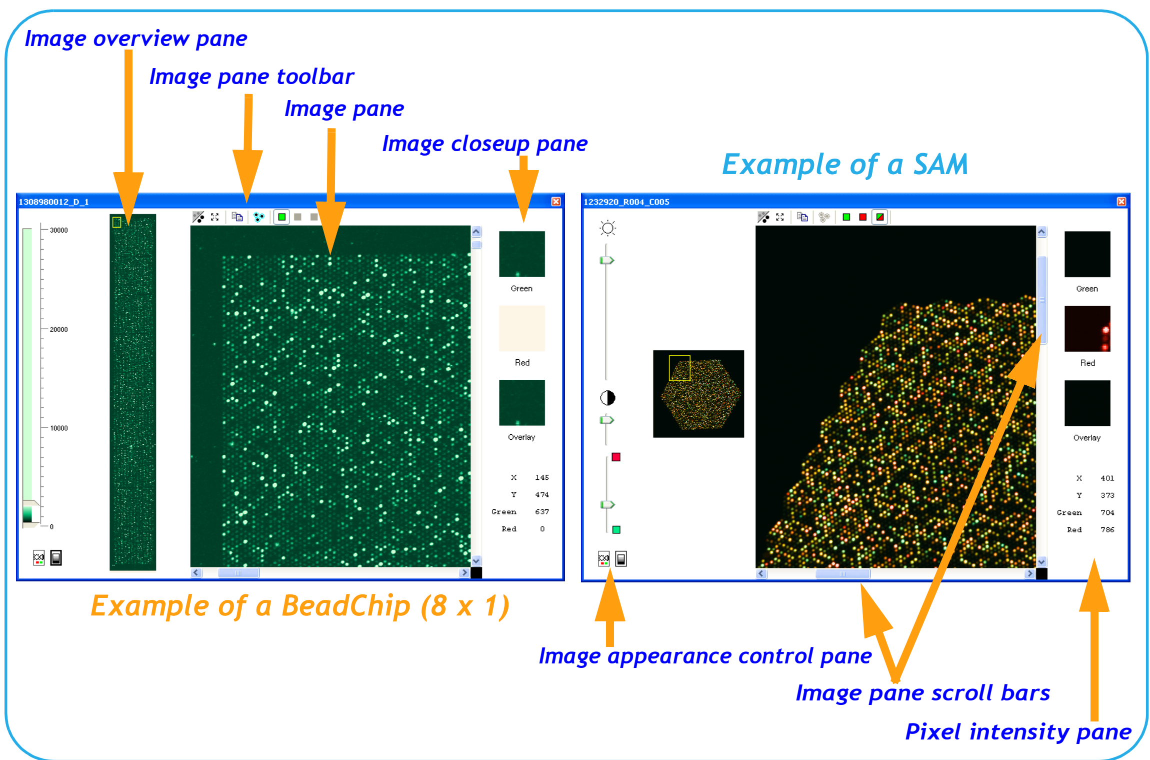


Figure 4-2 Sample Image

Table 4-1 Image Viewer Features

Feature	*Description*
Image overview pane	Displays a bird's-eye view of the sample image. Use the mouse to move the yellow box, which determines the Image pane's field of view.
Image pane tool bar	Click the **Auto Contrast** button to reset the image contrast to default value. Click the **Zoom Out** button to return to the default image view. Click the **Copy to Clipboard** button to save an image to the clipboard for pasting into another application. Click to verify successful registration during data extraction on the BeadArray Reader. See *Overlay Cores* on page 4-7 for details. Click to see the green channel only. Click to see the red channel only. Click to see both red and green channels.
Image pane	Allows detailed inspection of the sample image. Use the mouse slide wheel to control the zoom level. If your mouse lacks a wheel, zoom into a region by pressing the shift key and the left mouse button at the same time, then dragging to select the zoom area and releasing the mouse button and shift key. To zoom back out, click the **Zoom Out** button on the Image pane tool bar.

Table 4-1 Image Viewer Features (continued)

Feature	Description
Image closeup pane	Displays a closeup view of your image in the red and green color channels, and in a merged (overlay) view. The view region is determined by the location of your mouse pointer on the Image pane. *Note: The red color channel is disabled for monochrome direct hyb images.*
Image appearance control pane	Used to control image brightness, contrast, and color balance. See Figure 4-4 for details. *Note: These controls affect only the appearance of the image on the screen; they do not change the underlying image file.*
Image pane scroll bars	Allow you to change the viewing region in the Image pane.
Pixel intensity pane	Reports the X Y coordinates of your mouse pointer on the Image pane, along with the pixel intensity of that location.

Overlay Cores

Click **Overlay Cores** to display the Image pane as shown in Figure 4-3. This feature allows you to verify that registration succeeded during data extraction on the BeadArray Reader.

NOTE:
The Overlay Cores tool only functions when viewing either the red channel or the green channel. It will not function when viewing both channels simultaneously.

Zoom in on a corner of the image to see blue circles overlaying the scanned sample image spots. A successful registration is indicated when the boundary of the blue-circle grid coincides with the sample pixel boundary.

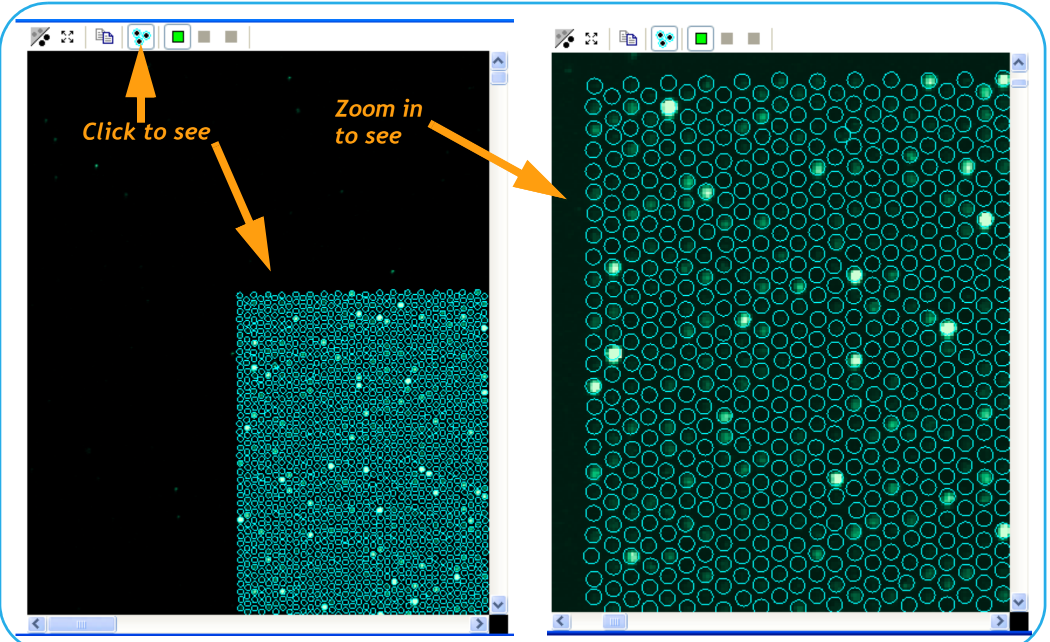


Figure 4-3 Overlay Cores Image

NOTE:
In rare cases, registration can fail. Contact Illumina Customer Solutions.

Image Appearance Control

You can control image appearance in two different modes:

- Brightness/Contrast mode
- Intensity Threshold mode

Use the mode selection buttons at the bottom of the pane to select the desired mode.

The Image appearance control pane components are described in Figure 4-4.

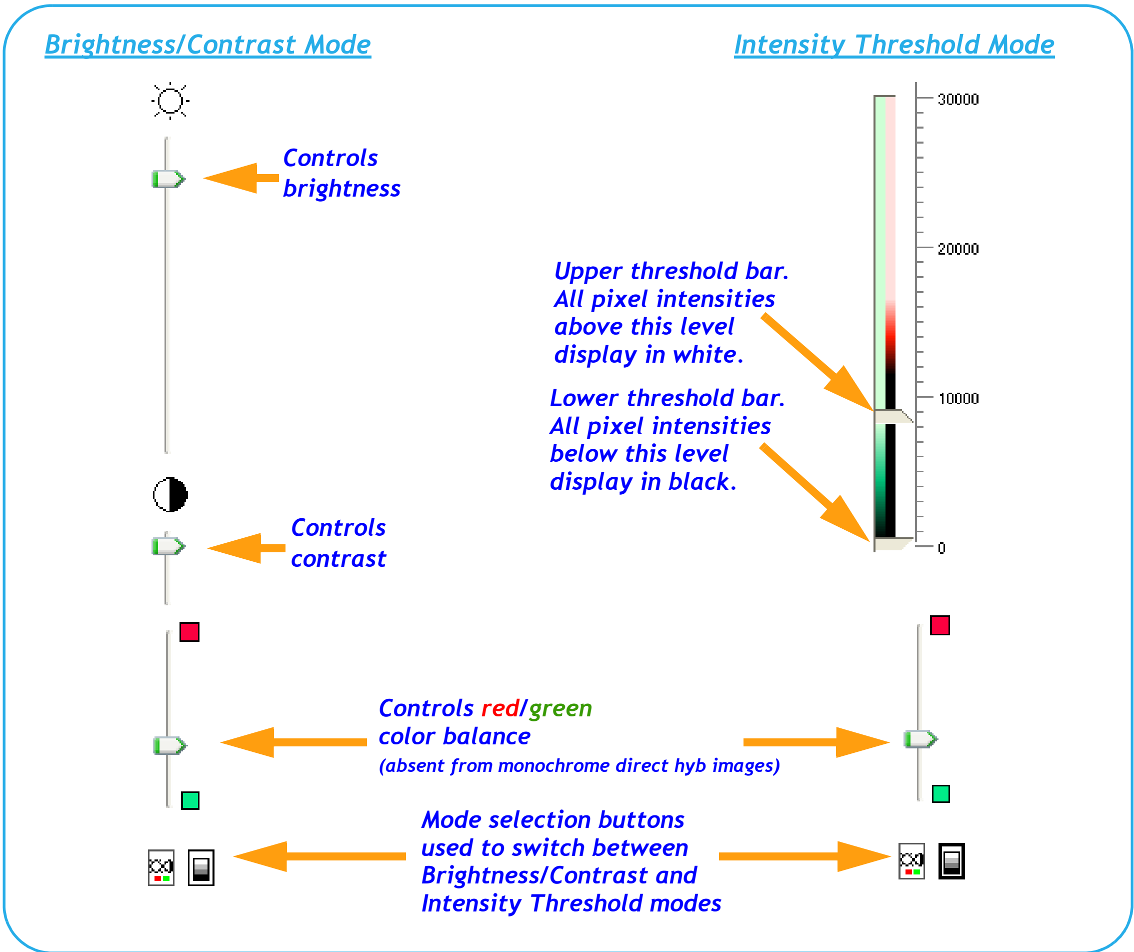


Figure 4-4 Image Control Pane Detail

NOTES

NOTES

Chapter 5

Screen Elements

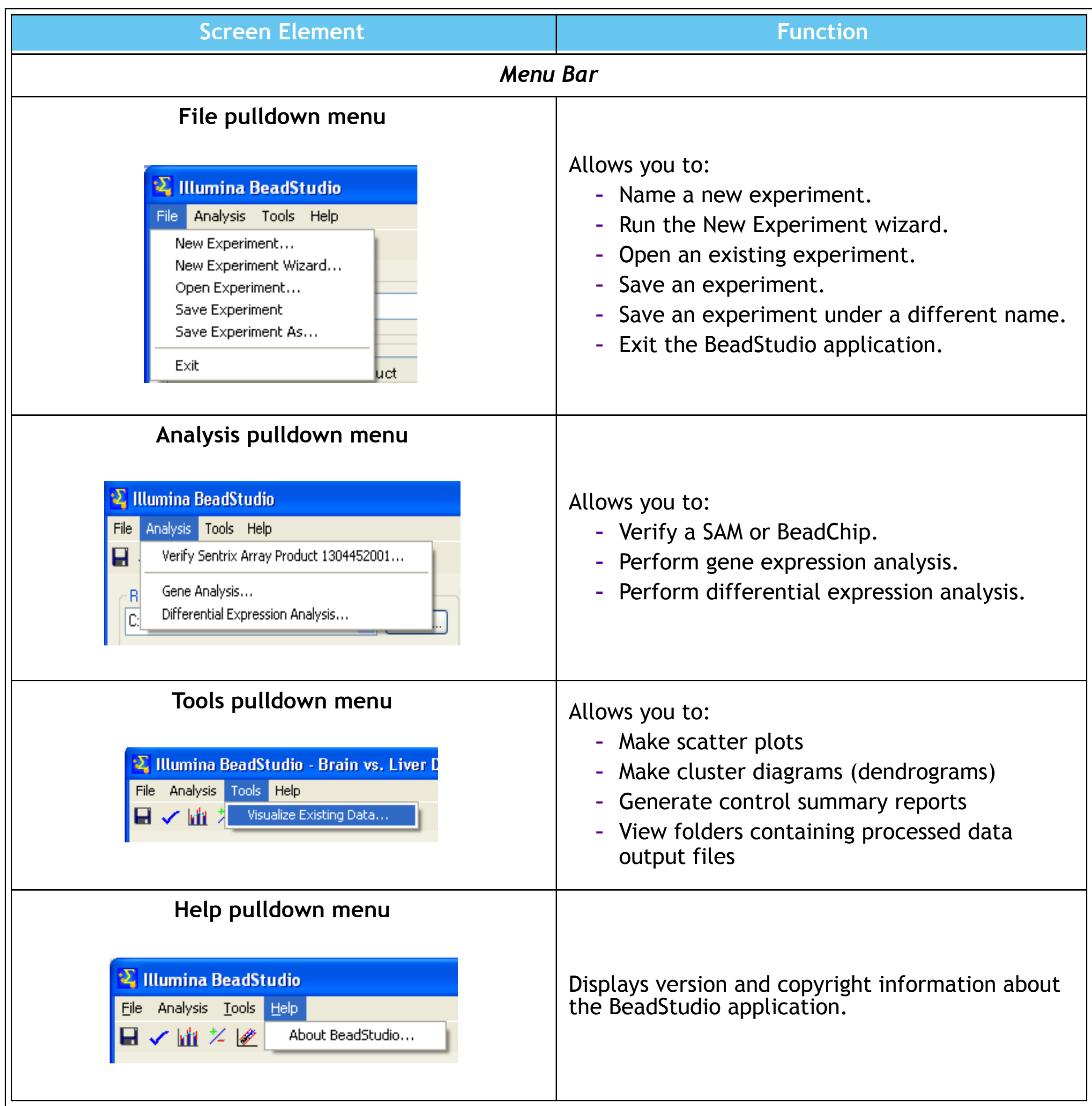
Topics

- BeadStudio User Interface Screen Elements *5-2*
- BeadStudio Example Screens *5-7*

BeadStudio User Interface Screen Elements

Refer to Table 5-1 for brief descriptions of the menus, buttons, fields, and panes of the BeadStudio application screens.

Table 5-1 BeadStudio User Interface Screen Elements



Screen Element	Function
Menu Bar	
File pulldown menu	Allows you to: - Name a new experiment. - Run the New Experiment wizard. - Open an existing experiment. - Save an experiment. - Save an experiment under a different name. - Exit the BeadStudio application.
Analysis pulldown menu	Allows you to: - Verify a SAM or BeadChip. - Perform gene expression analysis. - Perform differential expression analysis.
Tools pulldown menu	Allows you to: - Make scatter plots - Make cluster diagrams (dendrograms) - Generate control summary reports - View folders containing processed data output files
Help pulldown menu	Displays version and copyright information about the BeadStudio application.

Table 5-1 BeadStudio User Interface Screen Elements (continued)

Screen Element	Function
Main Toolbar Buttons	
Save experiment button	Click to save an experiment to the directory of your choice.
Verify Matrix/BeadChips button	Click to verify the uniformity of selected sample intensities.
Gene analysis button	Click to perform gene analysis for a default experiment.
Diff analysis button	Click to perform differential gene analysis for a default experiment.
Start scatter plot button	Click to create a scatter plot.
Start cluster analysis button	Click to create a dendrogram.

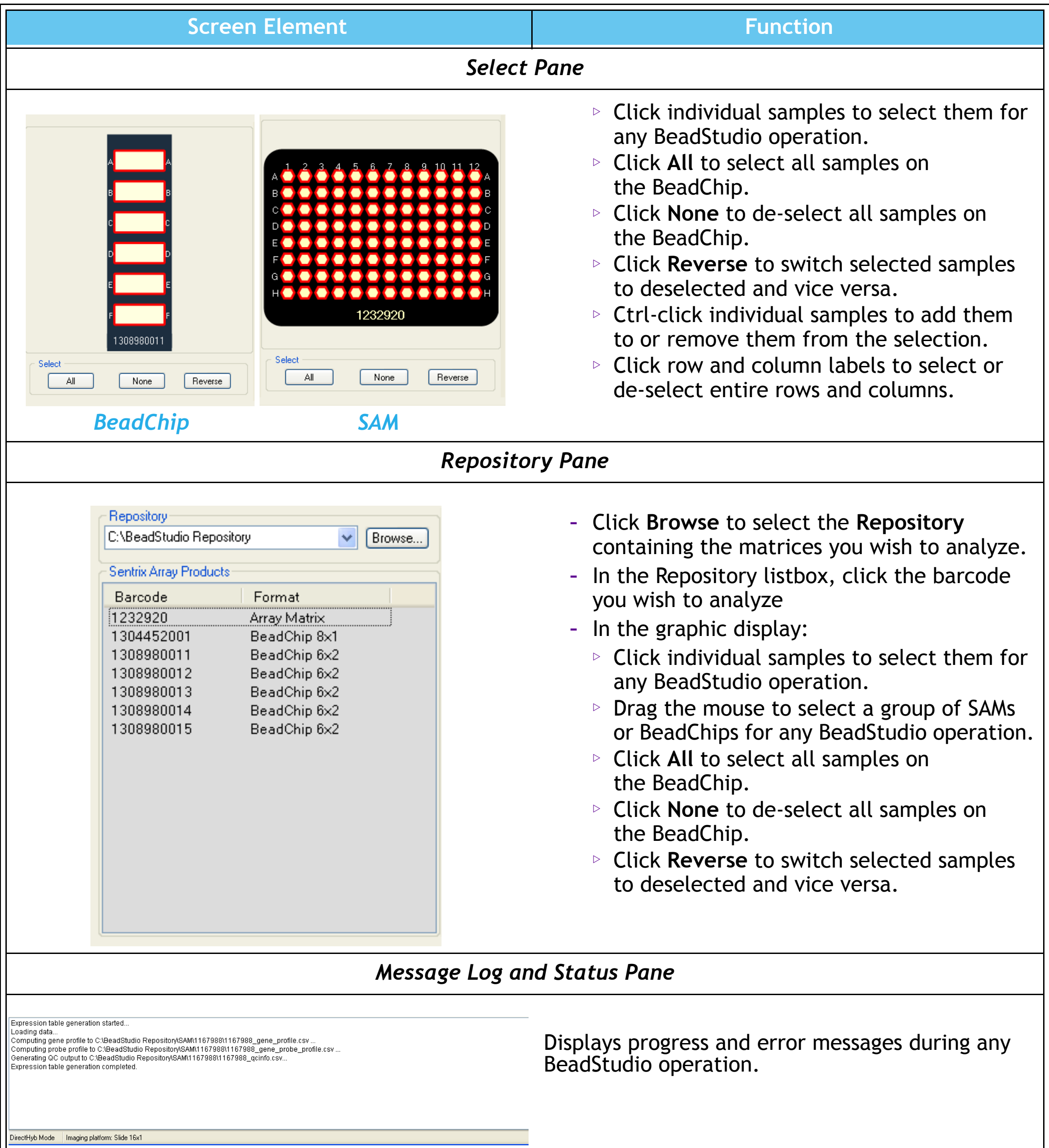
Table 5-1 BeadStudio User Interface Screen Elements (continued)

Screen Element	Function
Stop operation button	Click to stop any process you have asked BeadStudio to initiate.
Groups Toolbar Buttons	
New group... button	Click to add a new **Group** (name an experiment).
Add to Group button	Click to add selected samples to the selected group in the **Experiment** pane.
Create a group for each sample... button	Click to create a new experimental group for each selected sample.
Load from sample sheet... button	Click to create and populate new experimental groups using a sample sheet (see *Set Up & Apply Sample Sheet* on page 2-28).
Apply group layout... button	Click to create and populate new experimental groups using a group layout file (see *Set Up & Apply Group Layout Files* on page 2-25).

Table 5-1 BeadStudio User Interface Screen Elements (continued)

Screen Element	Function
Remove... button	Click to delete the selected groups or experiments within a group.
Clear all groups... button	Click to delete all experimental groups.
Groups Pane	
Groups Mock Hyb 1 1304452001_A 100% Liver 1 1304452001_B 95% L - 5 % B 1 1304452001_C 67% L - 33% B 1 1304452001_D 33% L - 67% B 1304452001_E 5% L - 95% B 1 1304452001_F 100% Brain 1 1304452001_G 100% Liver-rep 1 1304452001_H	Displays the groups and samples you have selected for the experiment. Contains a toolbar for creating, populating, and deleting groups.

Table 5-1 BeadStudio User Interface Screen Elements (continued)



Screen Element	Function
Select Pane	
BeadChip *SAM*	▹ Click individual samples to select them for any BeadStudio operation. ▹ Click **All** to select all samples on the BeadChip. ▹ Click **None** to de-select all samples on the BeadChip. ▹ Click **Reverse** to switch selected samples to deselected and vice versa. ▹ Ctrl-click individual samples to add them to or remove them from the selection. ▹ Click row and column labels to select or de-select entire rows and columns.
Repository Pane	
	- Click **Browse** to select the **Repository** containing the matrices you wish to analyze. - In the Repository listbox, click the barcode you wish to analyze - In the graphic display: ▹ Click individual samples to select them for any BeadStudio operation. ▹ Drag the mouse to select a group of SAMs or BeadChips for any BeadStudio operation. ▹ Click **All** to select all samples on the BeadChip. ▹ Click **None** to de-select all samples on the BeadChip. ▹ Click **Reverse** to switch selected samples to deselected and vice versa.
Message Log and Status Pane	
	Displays progress and error messages during any BeadStudio operation.

BeadStudio Example Screens

NOTE:

Refer to Table 5-2 and through 5-2 for color significance of displayed samples.

Table 5-2 Color Key for Displayed Samples

Sample Display Color	Definition
Red Outline	User has selected this sample for some action.
Tan	An extracted intensity file (.idat) exists for this sample. The sample is ready for gene expression analysis.
Speckled	No data exists for this sample (neither image nor extracted intensities).
Gray	The BeadStudio **Verify Matrix/BeadChip** task marked this sample as an outlier. This sample is suspected of having a registration problem or has significantly different intensity than the other samples (requires examination in Image Viewer).

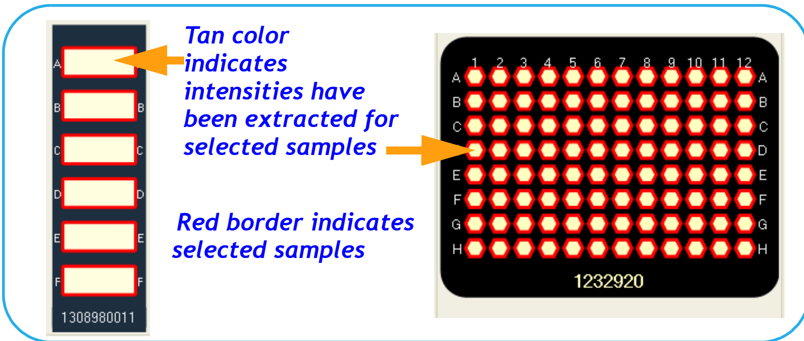


Figure 5-1 Extracted Intensity Samples Shown in Tan

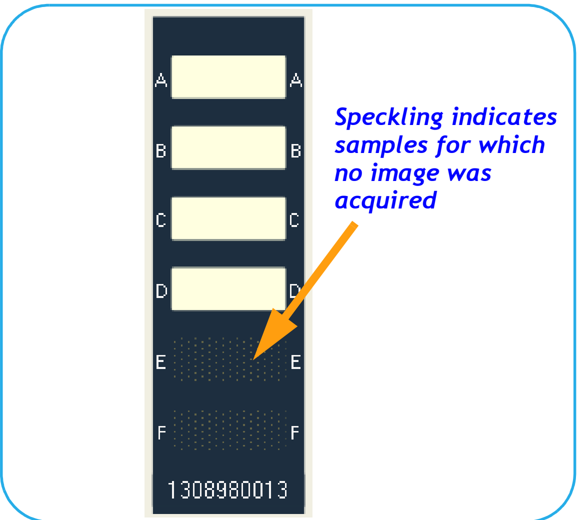


Figure 5-2 Unimaged Samples Shown as Speckled

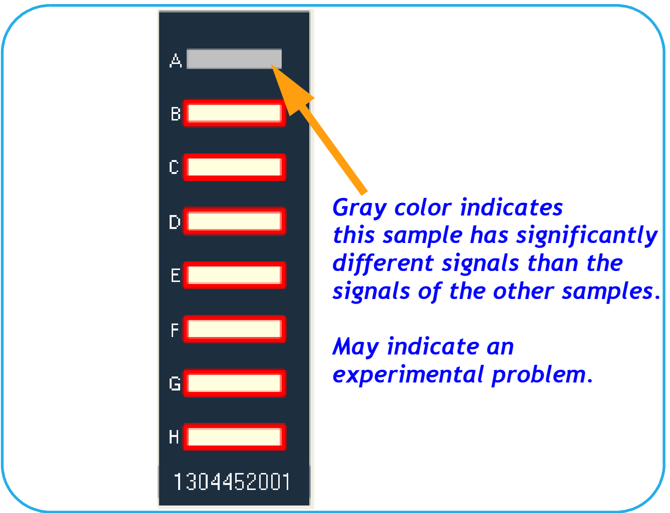


Figure 5-3 Extracted Intensities, Problem Samples in Gray

NOTES

NOTES

Chapter 6

Normalization & Differential Analysis Algorithms

Topics

- Introduction 6-2
- A Non-Mathematician's Guide 6-2
  - *Normalization Methods 6-2*
  - *Definitions 6-3*
  - *Background Method 6-3*
  - *Average Method 6-4*
  - *Cubic Spline Method 6-5*
  - *Hyb Controls Method 6-6*
  - *Rank-Invariant Method 6-7*
- Normalization Algorithms 6-8
  - *Background 6-8*
  - *Average 6-8*
  - *Cubic Spline 6-9*
  - *Hyb Controls 6-9*
  - *Rank Invariant 6-10*
- Differential Expression Algorithm 6-11
  - *Illumina Custom 6-11*
  - *Mann-Whitney 6-13*
  - *T-test 6-14*
- Detection Score 6-14
  - *Whole Genome BeadChips 6-14*
  - *Focused Array & DASL Products 6-15*

Introduction

This chapter describes the statistical algorithms used in expression analysis for Sentrix® arrays.

A Non-Mathematician's Guide

Normalization Methods

All methods of normalization aim to improve data by mathematically factoring out systematic errors among experimental groups so that their values can be compared. In the case of microarray experiments, systematic variation can result from variation in hybridization temperature, sample concentration, formamide concentration, etc. All forms of normalization achieve this result by making assumptions about the experimental samples and adjusting their values in a way that would factor out intensity changes arising from experimental variation without affecting changes based on true biological differences. The key to applying normalization effectively, therefore, is to understand the underlying assumptions of each method and deciding if they apply in the case of your experiment.

The sections below describe the normalization methods available in BeadStudio. For more rigorous mathematical descriptions, please see *Normalization Algorithms* on page 6-8. For the sake of simplicity, the explanations describe normalization as applied to two samples (A and B). The same principles apply when multiple samples are normalized together.

Definitions

When we speak of a sample, we refer to a single bundle on a Sentrix Array Matrix (SAM) or a single section of a Sentrix BeadChip. When we speak of a population of gene expression values, we refer to the set of all gene expression values received from a scan of a single sample. Therefore, normalization is a process by which two or more populations of gene expression values from two or more samples are adjusted for easier comparison. A scaling factor is a number by which values in one population are multiplied for the sake of normalization. For example, if a normalization technique multiplies all values in Sample B by 1.5 to normalize to Sample A, we say that a scaling factor of 1.5 was applied.

Background Method

This method subtracts a constant background value from each gene expression value acquired from a scanned sample. The background value is derived by averaging the signals of negative control beads built into the SAM or BeadChip. These beads or oligos contain sequences not expected to hybridize to most genomes and thus provide a measurement of non-specific hybridization, non-specific dye signal and scanner background. This method makes no biological assumptions about the samples and is thus safe to use when you have no expectations about the changes likely to exist between samples. Applying the technique allows for more quantitative assessments of fold-change differences, especially for genes with dim signals.

For the DASL Assay, the negative controls consist of oligos.

NOTE:

All other normalization methods described below apply background subtraction in addition to the other method-specific transformations.

Average Method

This method simply adjusts the intensities of two populations of gene expression values such that the means of the populations become equal. For example if the mean value for all genes in Sample A is 300 and the mean for Sample B is 100, all genes on Sample B will have their values scaled (multiplied) by a factor of 3 such that both populations now have a mean of 300. This method assumes that the mean expression levels of all genes should be roughly equal and similarly distributed. This assumption is generally true when samples contain large numbers of genes (such as with a whole-genome sample). However, the assumption breaks down when smaller numbers of genes are used or the samples are quite different. For example, if one had a focused sample containing a few hundred neurological genes and then used this sample to compare brain and liver sample, one would expect the brain sample to yield higher values for biological reasons; the method would therefore not apply. On the other hand, if one were comparing two brain samples, the assumption would probably apply and the method would be valid.

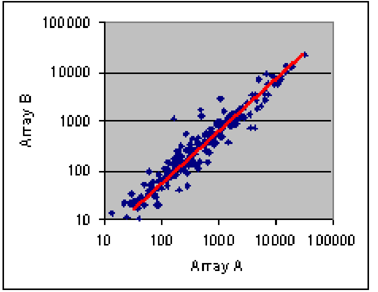
Cubic Spline Method

Cubic Spline normalization differs from all other methods described above in that it is non-linear. In other words, different scaling factors are applied to different parts of the population. The method first breaks the population of gene signals in each sample into a group of quantiles. If possible, the number of quantiles is chosen so that each interval contains 100 probe signals. However, the minimum number of quantiles is 15. For example, 3.3rd percentile, 10th percentile, 16.7th percentile and so on up to the 96.7th percentile. Then, for two samples to be normalized to each other, it scales the 3.3rd percentile of Sample B such that it is equal to the 3.3rd of Sample A, the 10th percentile of B to the 10th of A and so on for all quantiles. Genes whose values lie between quantiles are adjusted by interpolation of the neighboring quantiles.

The benefit of this method is that it can normalize between samples that show a non-linear relationship, such as can happen as a result of unequal sample labeling, different scan settings, etc.

The method is unnecessary when there is a linear relationship among the un-normalized signals of the populations, and in these cases cubic spline normalization should not be applied. To determine if your data has a linear relationship, you can use the Scatter Plot tool (described in Chapter 3, *Data Visualization*) to generate a scatter plot of un-normalized gene expression for the two samples.

If all gene signals from the samples are plotted against each other and show a generally linear relationship (such as in the left plot in Figure 6-1), the cubic spline normalization should not be applied.



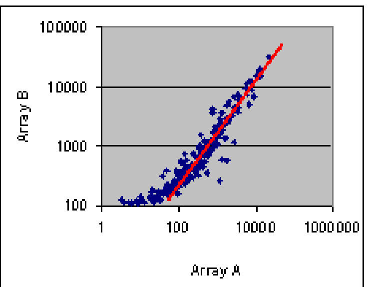


Figure 6-1 Sample A & Sample B

If there is a distortion in the linear relationship (as in the right plot in Figure 6-1), cubic spline may be applied and has the potential to correct the distortion.

Hyb Controls Method

This method works like the 'Average' method above, except, instead of using the signal of all genes to calculate the scaling factor between samples, it uses signals of positive control probes. These probes are included in every sample and hybridize to corresponding labeled oligonucleotides contained in our standard hybridization buffers. The advantage of this approach is that it allows a signal-based normalization of samples while making no biological assumptions about the similarity of the samples. However, due to differences between the dyes used in the control oligonucleotides and those used in the RNA labeling, as well as pipette errors, the hyb controls may be imperfect proxies for the genes in your sample. Also, the Hyb Controls method does not allow correction for differences arising from sample labeling. For these reasons, this method should be applied with caution.

Rank-Invariant Method

For most types of expression experiments, this is our most highly recommended normalization method. Like the Average method above, Rank-Invariant normalization uses a linear scaling of the populations being compared. However, unlike with averaging, the scaling factor is determined not by an average of all genes, but by only rank-invariant genes. 'Rank-invariant' genes are those whose expression values show a consistent order relative to other genes in the population. For example, a gene that is the 200th brightest gene in Sample A and 203rd in Sample B would be considered rank-invariant and would be used to arrive at the normalization factor; a gene that goes from 200th to 10,000th would not be rank-invariant and would not be used. This method is much more resistant to outliers than straight averaging is and generally gives better results. However, as with averaging, if samples are very different in their behaviors, the underlying assumption of rank-invariance (the existence of a subpopulation of genes whose expression is constant across samples showing consistent ranks) will not be true and the method should not be applied.

Due to the DASL Assay's oligo-directed nature, the assumption of similar behavior between samples is often not true. Although rank normalization is preferred for similar samples, the degree of similarity depends on: 1) gene expression in the samples analyzed; and 2) the genes chosen for the oligo pool. Illumina recommends examining un-normalized DASL data in scatter plots before choosing a normalization method for further analysis.

Normalization Algorithms

For all algorithms, normalization is computed with respect to a mathematically calculated “virtual” sample that represents averaged probe intensities across a group of samples. In the cases of spline and rank invariant normalizations, the virtual sample is computed based on the content of the reference group. If there is no reference group, the first group in the list of groups displayed in the Experiment pane is used for group analysis. For SAM/BeadChip analysis, the virtual sample is computed based on the content of the first alphanumeric entry in the upper-left area of the Matrix pane. For the hyb controls and average methods, all samples in the experiment are averaged to produce the virtual sample. A detailed description of normalization algorithms follows.

Background

The background value is derived by averaging the signals of built-in negative control Bead types. Outliers are removed using the median absolute deviation method.

Average

Sample intensities are simply scaled by a factor equal to the ratio of average intensity of virtual sample to the average intensity of the given sample. Background is subtracted prior to the scaling.

Cubic Spline

The method is similar to the one proposed by Workman et al.[1] The normalization uses quantiles of sample intensities to fit smoothing B-splines.

Let $q_i = \frac{(i-0.5)}{N}, i = 1, 2, \ldots, N$ be a vector of *N* quantiles ($N = max\left(15, \frac{N_{probes}}{100}\right)$). Here, N_{probes} is the number of probes represented on an sample.

For each sample, we compute its vector of quantile intensities. Similarly, we compute quantiles for the “virtual” averaged sample after background subtraction. Cubic B-spline is computed and used for interpolation. For points with intensities ranked outside the $[q_1, q_N]$ interval, we use linear extrapolation rather than spline to avoid nonlinear effects outside the region of interpolation.

Hyb Controls

Let *k = 1 ..., N* enumerate all samples used in the experiment. Then for sample *k*, normalization coefficients (a_k, b_k) are computed using iteratively re-weighted least-squares fit $y_v = a_k y_k + b_l$. Here, y_v, y_k are vectors of intensities of probes corresponding to hybridization controls on virtual and sample *k*, respectively. Tukey bisquare weight function with tuning constant 4.685 provides 95% efficiency when errors are normally distributed, while maintaining protection against outliers. Standard deviation of errors is estimated using median absolute deviation. Normalized intensities are computed with $y_k^{new} = \frac{y_k - b_k}{a_k}$, and then background is subtracted. For further information on the use of hyb controls, see the System Manual *System Controls* appendix for your specific product.

---

1. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 2002 Aug 30;3(9):research0048. PMID: 12225587 [PubMed - indexed for MEDLINE]

Rank Invariant

This method is exactly the same as *Hyb Controls* on page 6-9, except it uses a rank invariant set of probes between a given sample and a virtual sample instead of hybridization controls. The rank invariant set is found as follows: we start by considering probes with intensities ranked between LowRank=50th percentile and HighRank=90th percentile. If the probe's relative rank change $\frac{|r_x - r_v|}{r_v} \leq 0.05$, then the probe is considered to be rank invariant. If less than 2% of all probes in the region are identified as rank invariant, then LowRank is gradually decreased until it reaches 25th percentile.

Differential Expression Algorithm

All algorithms compare a group of samples (referred to as the condition group) to a reference group. The comparison is done using the following error models:

- Illumina custom
- Mann-Whitney
- T-test

Illumina Custom

This model assumes that target signal intensity (I) is normally distributed among replicates corresponding to some biological condition. The variation has three components: sequence specific biological variation (σ_{bio}), nonspecific biological variation (σ_{neg}), and technical error (σ_{tech}).

$$I = N(\mu, \sigma)$$
$$\sigma = \sqrt{\sigma_{tech}^2 + \sigma_{neg}^2 + \sigma_{bio}^2}$$
$$\sigma_{tech} = a + b < I >$$

Variation of nonspecific signal σ_{neg} is estimated from the signal of negative control sequences (using median absolute deviation). For σ_{tech}, we estimate two sets of parameters (a_{ref}, b_{ref}) and (a_{cond}, b_{cond}) for reference and condition groups respectively.

We estimate σ_{tech} using iterative robust least squares fit which reduces influence of highly variable genes. This implicitly assumes that the majority of genes do not have high biological variation among replicates. When this assumption does not hold we overestimate technical error by some averaged biological variation.

When groups contain biological replicates, we produce p-values using the following approach:

$$\sigma_{ref} = \max(s_{ref},\ a_{ref} + b_{ref} I_{ref})$$
$$\sigma_{cond} = \max(s_{cond}, a_{cond} + b_{cond} I_{cond})$$

$$p = Z\left(\frac{|I_{cond} - I_{ref}|}{\sqrt{\frac{\sigma_{ref}^2 + \sigma_{neg(ref)}^2}{N_{ref}} + \frac{\sigma_{cond}^2 + \sigma_{neg(cond)}^2}{N_{cond}}}} \right)$$

where S_{ref} and S_{cond} are standard deviations of probe signals.

NOTE:

N_{ref} and N_{cond} denote the number of samples in the reference and condition groups respectively.

We consider that standard deviations exceeding σ_{tech} reflects biological variation. However, we assume that estimates smaller than σ_{tech} are caused by random errors. Therefore, we use the larger of two estimates. Usage of σ_{neg} provides regularization for low abundance targets. Z is two-sided tail probability of standard normal distribution.

When reference and conditions groups contain one sample each, we can neither estimate sequence specific biological variation nor sample processing variation. Instead, we can only assess σ using bead type variation. Therefore, we penalize for that by a factor of 2.5 applied to parameter b. This factor was determined empirically from examination of real sample data.

$$p = Z\left(\frac{|I_{cond} - I_{ref}|}{\sqrt{(a_{ref} + 2.5 b_{ref} I_{ref})^2 + \sigma_{neg(ref)}^2 + (a_{cond} + 2.5 b_{cond} I_{cond})^2 + \sigma_{neg(cond)}^2}} \right)$$

In DASL mode, this factor is 3.

A DiffScore for a probe is computed as:

$$DiffScore = 10\,\mathrm{sgn}(\mu_{cond} - \mu_{ref})\log_{10}(p)$$

For the gene, DiffScores of corresponding probes are averaged. In addition, concordance between probes is reported.

In DASL, p-values are generated for red and green channels independently. These are averaged and the final p-value is generated from the distribution of the average of two independent uniform (0, 1) variables. If direction of intensity change is different for red and green signals, then the larger of p-values is replaced by 1 - p-value prior to averaging.

Concordance is defined as $\frac{|n_u - n_d|}{|n_u + n_d|}$ where n_u is the number of probes showing upregulated signal and n_d is the number of probes showing downregulated signal.

Mann-Whitney

This implementation produces exact p-value if:

$$\min(N_{ref}, N_{cond}) < 3$$

OR

$$\min(N_{ref}, N_{cond}) < 9 \quad AND \quad \max(N_{ref}, N_{cond}) < 13$$

Otherwise, normal approximation with continuity correction is used. Differential scores are computed as described for the Illumina Custom model (page 6-11).

T-test

When either the reference group or a condition group contains at least two samples, variance is estimated across replicate samples. Otherwise, variance is estimated from bead-to-bead variation*. We use t-test with the assumption of equal variance.

* *Variance computed from bead-to-bead variation may significantly underestimate total variance. We recommend using Illumina Custom model in this case.*

Differential scores are computed the same way as described for the Illumina Custom model (see page 6-7).

Detection Score

Detection scores are computed using negative control signals. Because Illumina's whole genome BeadChips contain large numbers of negative controls (1,000 - 2000), while its focused array and DASL products contain fewer negative controls, different algorithms are used for each type of product.

Whole Genome BeadChips

For whole genome BeadChips, the detection algorithm uses a large number of negative control probes.

Instead of using parametric assumptions, gene signals are ranked relative to the distribution of signals of the negative controls.

DetectionScore = R / N, where R is the rank of the gene signal relative to negative controls and N is number of negative controls. For groups containing multiple samples, the following modification is used. Let m be the number of samples in the group. On the ith sample, the gth gene signal is converted to a Z value and the average Z value across all m samples is computed.

$$Z_g = \frac{1}{m}\sum_i Z_g^i$$

$$Z_g^i = \frac{I - \mu^i_{neg}}{\sigma^i_{neg}}$$

Here μ^i_{neg} and σ^i_{neg} are the mean and standard deviation of signals of the negative controls on the ith sample. I is the signal from gene g. The same transformation that is applied to I is also applied to the signals of negative controls. Detection Scores are computed based on the rank of the Z value of a gene relative to the Z values of the negative controls.

Focused Array & DASL Products

Since these products typically contain small numbers of negative controls (20 - 40), their signals (with outliers removed using median absolute deviation) are modeled by normal distribution. The detection score for the probe with intensity I_{probe} is given by:

$$1 - Z\left(\frac{|I_{probe} - \mu_{neg}|}{\sigma_{neg}}\right)$$

Here, μ_{neg} is the average intensity of negative controls and σ_{neg} is the standard deviation of their signals. Z is the one-sided tail probability of standard normal distribution. For the gene represented by N probes we use:

$$1 - Z\left(\sqrt{N}\frac{|I_{gene} - \mu_{neg}|}{\sigma_{neg}}\right)$$

When experimental group contains M replicate samples, the average Z value of $Z_1,\ldots,Z_M$, computed for each sample independently, is assumed to follow a normal distribution

$$N\left(0, \sqrt{\frac{1 + (M-1)r}{M}}\right)$$

where r is the average correlation coefficient of signals of negative controls.

Averaging is done across all pairs of different samples.

In DASL mode, detection p-values (1 - DetectionScore) in red and green channels are computed independently. Their average is assumed to follow distribution of the average of two independent variables distributed uniformly on the interval (0,1). The p-value is generated from that distribution and converted to DetectionScore as 1-pvalue.

NOTES

NOTES