User’s Guide I

Quick Start, Introduction, Importing Microarray Data, Data Utilities, and Graphics

GESS
Gene Expression Statistical System

Published by
NCSS
Christopher J. Hintze, Ph.D.
Kaysville, Utah

GESS User’s Guide I

Direct inquiries to:

NCSS
329 North 1000 East
Kaysville, Utah 84037
Phone (801) 546-0445
Fax (801) 546-3907
Email: support@ncss.com

Warning:

This software and manual are both protected by U.S. Copyright Law (Title 17 United States Code). Unauthorized reproduction and/or sales may result in imprisonment of up to one year and fines of up to $10,000 (17 USC 506). Copyright infringers may also be subject to civil liability.

About This Manual

Congratulations on your purchase of the ***GESS*** package! ***GESS*** offers:

- Easy file specification.
- A comprehensive list of accurate microarray analysis routines that are quick to learn and easy to use.
- Straightforward procedures for creating paper printouts and file copies of both the numerical and graphical reports.

Our goal is that with the help of these user's guides, you will be up and running on ***GESS*** quickly. After reading the introductory chapters, you will only need to refer to the chapters corresponding to the procedures you want to use. The discussion of each procedure includes one or more tutorials that will take you step-by-step through the tasks necessary to run the procedure.

I believe you will find that these user’s guides provide a quick, easy, efficient, and effective way for first-time ***GESS*** users to get up and running.

I look forward to any suggestions you have to improve the usefulness of this manual and/or the ***GESS*** system. Meanwhile, good computing!

Christopher Hintze, Author

GESS License Agreement

Important: The enclosed Gene Expression Statistical System software program (GESS) is licensed by NCSS, LLC to customers for their use only on the terms set forth below. Purchasing the system indicates your acceptance of these terms.

1. **LICENSE.** NCSS, LLC hereby agrees to grant you a non-exclusive license to use the accompanying GESS program subject to the terms and restrictions set forth in this License Agreement.

2. **COPYRIGHT.** GESS and its documentation are copyrighted. You may not copy or otherwise reproduce any part of GESS or its documentation, except that you may load GESS onto a computer as an essential step in executing it on the computer and make backup copies for your use on the same computer.

3. **BACKUP POLICY.** GESS may be backed up by you for your use on the same machine for which GESS was purchased.

4. **RESTRICTIONS ON USE AND TRANSFER.** The original and any backup copies of GESS and its documentation are to be used only in connection with a single user. This user may load GESS onto several machines for his/her convenience (such as a desktop and laptop computer), but only for use by the licensee. You may physically transfer GESS from one computer to another, provided that GESS is used in connection with only one user. You may not distribute copies of GESS or its documentation to others. You may transfer this license together with the original and all backup copies of GESS and its documentation, provided that the transferee agrees to be bound by the terms of this License Agreement. GESS licenses may not be transferred more frequently than once in twelve months. Neither GESS nor its documentation may be modified or translated without written permission from NCSS, LLC.

*You may not use, copy, modify, or transfer **GESS**, or any copy, modification, or merged portion, in whole or in part, except as expressly provided for in this license.*

5. **NO WARRANTY OF PERFORMANCE.** NCSS, LLC does not and cannot warrant the performance or results that may be obtained by using GESS. Accordingly, GESS and its documentation are licensed "as is" without warranty as to their performance, merchantability, or fitness for any particular purpose. The entire risk as to the results and performance of GESS is assumed by you. Should GESS prove defective, you (and not NCSS, LLC or its dealer) assume the entire cost of all necessary servicing, repair, or correction.

6. **LIMITED WARRANTY ON CD.** To the original licensee only, NCSS, LLC warrants the medium on which GESS is recorded to be free from defects in materials and faulty workmanship under normal use and service for a period of ninety days from the date GESS is delivered. If, during this ninety-day period, a defect in a cd should occur, the cd may be returned to NCSS, LLC at its address, or to the dealer from which GESS was purchased, and NCSS, LLC will replace the cd without charge to you, provided that you have sent a copy of your receipt for GESS. Your sole and exclusive remedy in the event of a defect is expressly limited to the replacement of the cd as provided above.

Any implied warranties of merchantability and fitness for a particular purpose are limited in duration to a period of ninety (90) days from the date of delivery. If the failure of a cd has resulted from accident, abuse, or misapplication of the cd, NCSS, LLC shall have no responsibility to replace the cd under the terms of this limited warranty. This limited warranty gives you specific legal rights, and you may also have other rights, which vary from state to state.

7. **LIMITATION OF LIABILITY.** Neither NCSS, LLC nor anyone else who has been involved in the creation, production, or delivery of GESS shall be liable for any direct, incidental, or consequential damages, such as, but not limited to, loss of anticipated profits or benefits, resulting from the use of GESS or arising out of any breach of any warranty. Some states do not allow the exclusion or limitation of direct, incidental, or consequential damages, so the above limitation may not apply to you.

8. **TERM.** The license is effective until terminated. You may terminate it at any time by destroying GESS and documentation together with all copies, modifications, and merged portions in any form. It will also terminate if you fail to comply with any term or condition of this License Agreement. You agree upon such termination to destroy GESS and documentation together with all copies, modifications, and merged portions in any form.

9. **YOUR USE OF GESS ACKNOWLEDGES** that you have read this customer license agreement and agree to its terms. You further agree that the license agreement is the complete and exclusive statement of the agreement between us and supersedes any proposal or prior agreement, oral or written, and any other communications between us relating to the subject matter of this agreement.

NCSS, LLC, Kaysville, Utah

Preface

GESS (**G**ene **E**xpression **S**tatistical **S**ystem) is an advanced, easy-to-use statistical analysis software package for microarray data. NCSS, LLC maintains a website at WWW.NCSS.COM where we make the latest edition of ***GESS*** available for free downloading. The software is password protected, so only users with valid serial numbers may use this downloaded edition. We hope that you will download the latest edition routinely and thus avoid any bugs that have been corrected since you purchased your copy.

We believe ***GESS*** to be an accurate, exciting, easy-to-use program. If you find any portion which you feel needs to be changed, please let us know. Also, we openly welcome suggestions for additions and enhancements.

User’s Guide I

Table of Contents

Quick Start

Introduction

Import Microarray Data

Data Utilities

Graphics

References and Indices

User’s Guide II
Table of Contents

Fold-Change Analysis

T-Tests

Analysis of Variance

Regression

Multivariate Analysis

References and Indices

Chapter 1

Installation and Basics

Before You Install

1. Check System Requirements

GESS runs on 32-bit and 64-bit Windows systems. This includes Windows ME, Windows NT 4.0, Windows 2000, Windows XP, and Windows Vista. The recommended minimum system is a Windows XP or Vista-compatible PC.

GESS takes up about 53 MB of disk space. Once installed, ***GESS*** also requires about 20 MB of temporary disk space while it is running. ***NCSS*** must be installed to run ***GESS***, but a license for ***NCSS*** is not required. ***NCSS*** requires about 120 MB of disk space.

2. Find a Home for GESS

Before you start installing, decide on a directory where you want to install ***NCSS*** and ***GESS***. By default, the setup program will install ***NCSS*** and ***GESS*** in the *C:\Program Files\NCSS\NCSS 2007* directory. You may change this during the installation, but not after. The example data, template, and macro files will be placed in your personal documents folder (usually *C:\…\[My] Documents\NCSS\NCSS 2007*) in appropriate subdirectories. The program will save all procedure templates and macros to these folders while the program is running.

3. If You Already Own NCSS or GESS

GESS and ***NCSS*** have been combined into one physical program. Access to each program is controlled by separate serial numbers. If you have a serial number for ***GESS***, but not for ***NCSS***, ***NCSS*** will work as trial versions for 7 days from the time the first procedure is accessed. Some of the ***NCSS*** routines (Scatter Plots, Box Plots, Histograms, and Data Report) are included with your purchase of ***GESS***.

If ***NCSS*** or ***GESS*** is already installed on your system, instruct the installation program to place this new version in a new folder (e.g. *C:\Program Files\NCSS\NCSS 2007*). All appropriate files will be copied from your old ***NCSS*** directory or replaced by updated files.

What Install Does

The installation procedure creates the necessary folders and copies the ***NCSS/GESS*** program from the installation file, called *NCSS2007SETUP.EXE*, to those folders. The files in

NCSS2007SETUP.EXE are compressed, so the installation program decompresses these files as it copies them to your hard disk. The following folders are created during installation:

C:\Program Files\NCSS\NCSS 2007 (or your substitute folder) contains most of the program files.

C:\Program Files\NCSS\NCSS 2007\Data\GESS contains data files required for ***GESS*** tutorials.

C:\Program Files\NCSS\NCSS 2007\Pdf\GESS contains printable copies of the documentation in PDF format.

C:\Program Files\NCSS\NCSS 2007\Sts contains all labels, text, and online messages.

C:\…\[My] Documents\NCSS\NCSS 2007\Data\GESS contains the database files used by the tutorials. We recommend creating a sub-folder of this folder to contain the data for each project you work on.

C:\…\[My] Documents\NCSS\NCSS 2007\Junk contains temporary files used by the program while it is running. Under normal operation, ***GESS*** will automatically delete temporary files. After finishing ***GESS***, you can delete any files left in this folder (but not the folder itself).

C:\…\[My] Documents\NCSS\NCSS 2007\Macros contains saved macros.

C:\…\[My] Documents\NCSS\NCSS 2007\Report is the default folder in which to save your output. You can save the reports to any folder you wish.

C:\…\[My] Documents\NCSS\NCSS 2007\Settings contains the files used to store your template files. These files are used by the ***GESS*** template system, which is described in a later chapter.

Installing NCSS and GESS

This section gives instructions for installing ***NCSS*** and ***GESS*** on your computer system. You must use the ***NCSS/GESS*** setup program to install ***GESS***. The files are compressed, so you cannot simply copy the files to your hard drive.

Follow these basic steps to install ***NCSS*** and ***GESS*** on your computer system.

Step	Notes
1.	Make sure that you are using a 32- or 64-bit version of windows such as Windows Me, Windows NT 4.0, Windows 2000, Windows XP, or Windows Vista.
2.	If you are installing from a CD, insert the CD in the CD drive. The installation program should start automatically. If it does not, on the Start menu, select the Run command. Enter *D:\NCSS\NCSS2007Setup*. You may have to substitute the appropriate letter for your CD drive if it is not *D*. If you are installing from a download, simply run the downloaded file (*NCSS2007Setup.exe*).
3.	Once the setup starts, follow the instructions on the screen. ***NCSS*** and ***GESS*** will be installed in the drive and folder you designate.

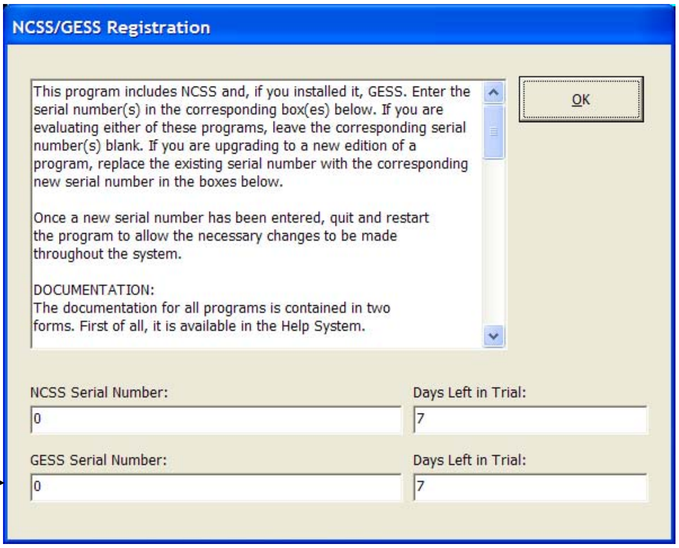
If Something Goes Wrong During Installation

The installation procedure is automatic. If something goes wrong during installation, delete the *C:\Program Files\NCSS\NCSS 2007* directory and start the installation process at the beginning. If trouble persists, contact our technical support staff as indicated below.

Starting GESS

GESS may be started using your keyboard or your mouse using the same techniques that you use to start any other Windows application. You can start ***GESS*** by selecting **NCSS 2007** from your Start menu using standard mouse or keyboard operations.

The first time you run ***GESS***, enter your serial number in the pop-up window that appears when the program begins. If you also have a serial number for ***NCSS***, enter it in the corresponding window. If you do not enter a serial number, ***NCSS*** and/or ***GESS*** will enter trial mode and you will have 7 days to evaluate ***NCSS*** and ***GESS***. When in trial mode, ***NCSS*** and ***GESS*** are fully-functional but limited to 100 rows of data.



NCSS/GESS Registration

This program includes NCSS and, if you installed it, GESS. Enter the serial number(s) in the corresponding box(es) below. If you are evaluating either of these programs, leave the corresponding serial number(s) blank. If you are upgrading to a new edition of a program, replace the existing serial number with the corresponding new serial number in the boxes below.

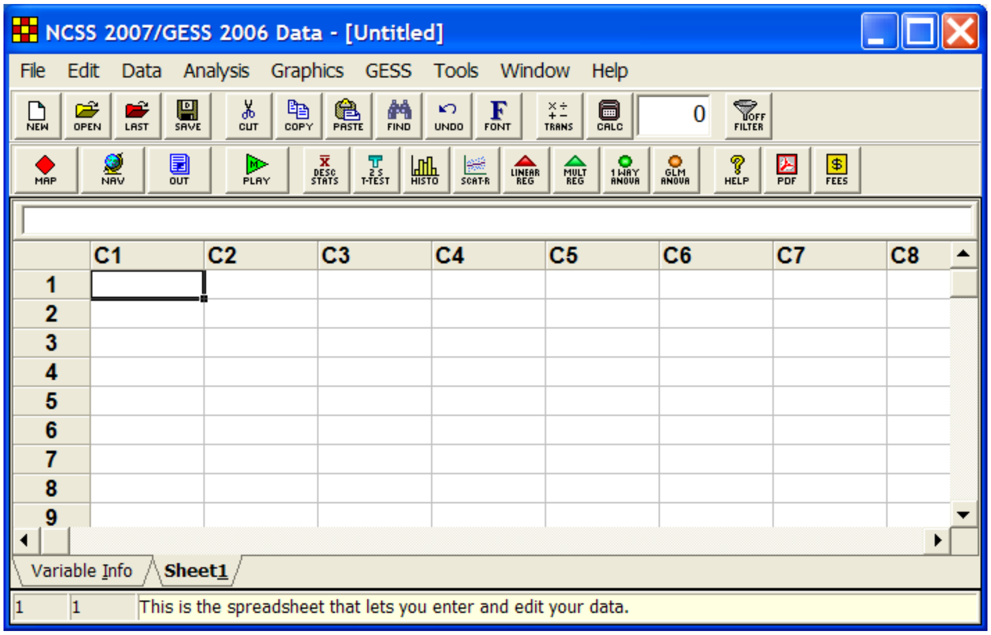
Once a new serial number has been entered, quit and restart the program to allow the necessary changes to be made throughout the system.

DOCUMENTATION:
The documentation for all programs is contained in two forms. First of all, it is available in the Help System.

NCSS Serial Number: 0
Days Left in Trial: 7
GESS Serial Number: 0
Days Left in Trial: 7

Enter your ***GESS*** serial number here.

The ***NCSS/GESS Data*** window will appear.



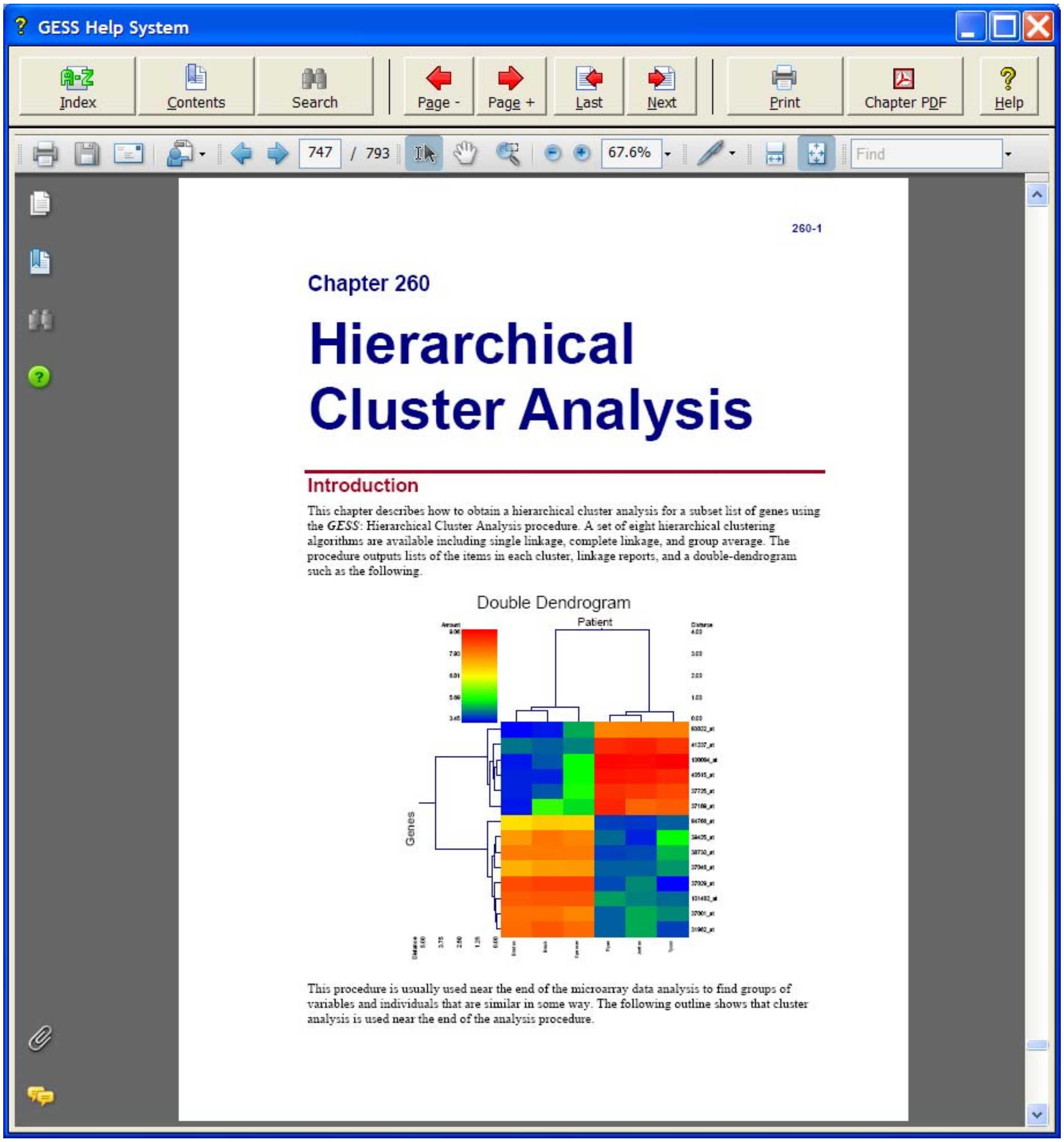
NCSS 2007/GESS 2006 Data - [Untitled]

1 1 This is the spreadsheet that lets you enter and edit your data.

Obtaining Help

Help System

To help you learn and use ***GESS*** efficiently, the material in this manual is included in the ***GESS*** Help System. The Help System is started from the Help menu or by clicking on the yellow '?' icon on the right side of the toolbar. ***GESS*** updates, available for download at www.ncss.com, may contain adjustments or improvements of the ***GESS*** Help System. Adobe Acrobat or Adobe Reader version 7 or 8 is required to view the help system. You can download Adobe Reader 8 for free by going to www.adobe.com. Adobe Reader 8 can also be installed from the *Utilities* folder on your ***GESS*** installation CD.



260-1

Chapter 260

Hierarchical Cluster Analysis

Introduction

This chapter describes how to obtain a hierarchical cluster analysis for a subset list of genes using the *GESS*: Hierarchical Cluster Analysis procedure. A set of eight hierarchical clustering algorithms are available including single linkage, complete linkage, and group average. The procedure outputs lists of the items in each cluster, linkage reports, and a double-dendrogram such as the following.

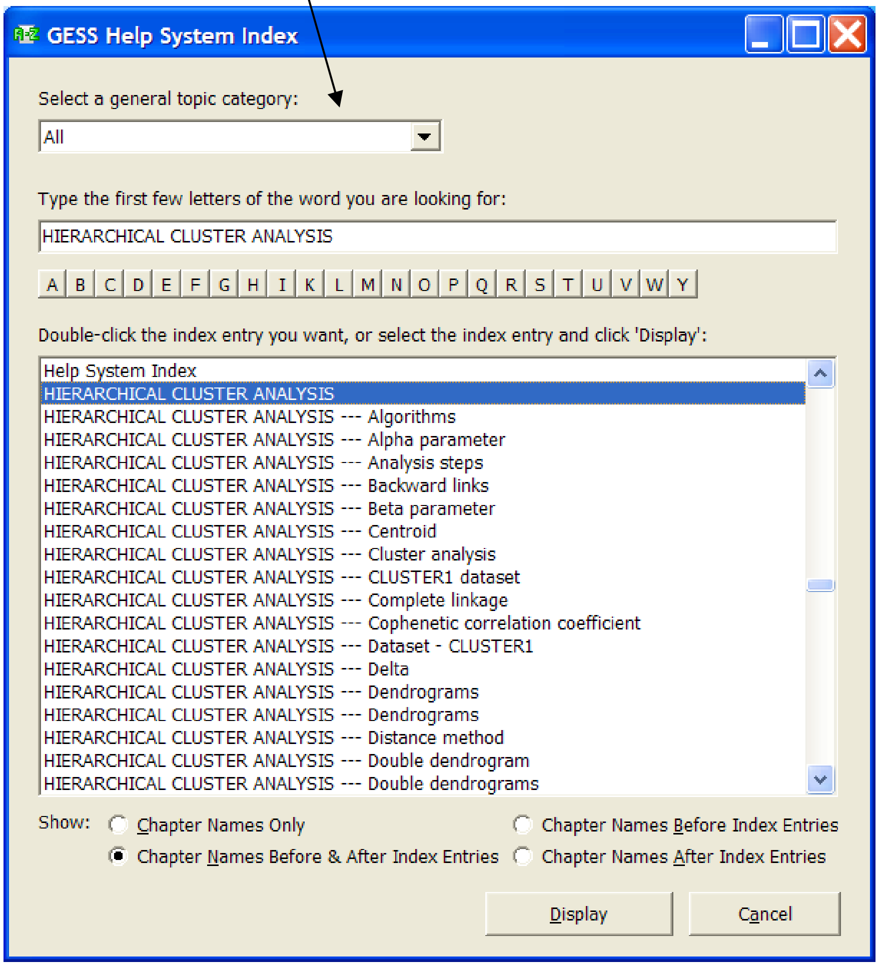
This procedure is usually used near the end of the microarray data analysis to find groups of variables and individuals that are similar in some way. The following outline shows that cluster analysis is used near the end of the analysis procedure.

Navigating the Help System

There are a few key features of our help system that will let you use the help system more efficiently. We will now explain each of these features.

Index Window

The Index Window can be launched at any time by clicking on the Index button on the ***GESS*** Help System display window. The index allows you to quickly locate keywords and/or statistical topics. You can narrow the list of index entries displayed by selecting a specific topic category in the uppermost dropdown box.



Select a general topic category: All

Type the first few letters of the word you are looking for: HIERARCHICAL CLUSTER ANALYSIS

Double-click the index entry you want, or select the index entry and click 'Display':

Help System Index
HIERARCHICAL CLUSTER ANALYSIS
HIERARCHICAL CLUSTER ANALYSIS --- Algorithms
HIERARCHICAL CLUSTER ANALYSIS --- Alpha parameter
HIERARCHICAL CLUSTER ANALYSIS --- Analysis steps
HIERARCHICAL CLUSTER ANALYSIS --- Backward links
HIERARCHICAL CLUSTER ANALYSIS --- Beta parameter
HIERARCHICAL CLUSTER ANALYSIS --- Centroid
HIERARCHICAL CLUSTER ANALYSIS --- Cluster analysis
HIERARCHICAL CLUSTER ANALYSIS --- CLUSTER1 dataset
HIERARCHICAL CLUSTER ANALYSIS --- Complete linkage
HIERARCHICAL CLUSTER ANALYSIS --- Cophenetic correlation coefficient
HIERARCHICAL CLUSTER ANALYSIS --- Dataset - CLUSTER1
HIERARCHICAL CLUSTER ANALYSIS --- Delta
HIERARCHICAL CLUSTER ANALYSIS --- Dendrograms
HIERARCHICAL CLUSTER ANALYSIS --- Dendrograms
HIERARCHICAL CLUSTER ANALYSIS --- Distance method
HIERARCHICAL CLUSTER ANALYSIS --- Double dendrogram
HIERARCHICAL CLUSTER ANALYSIS --- Double dendrograms

Show: ○ Chapter Names Only ○ Chapter Names Before Index Entries
● Chapter Names Before & After Index Entries ○ Chapter Names After Index Entries

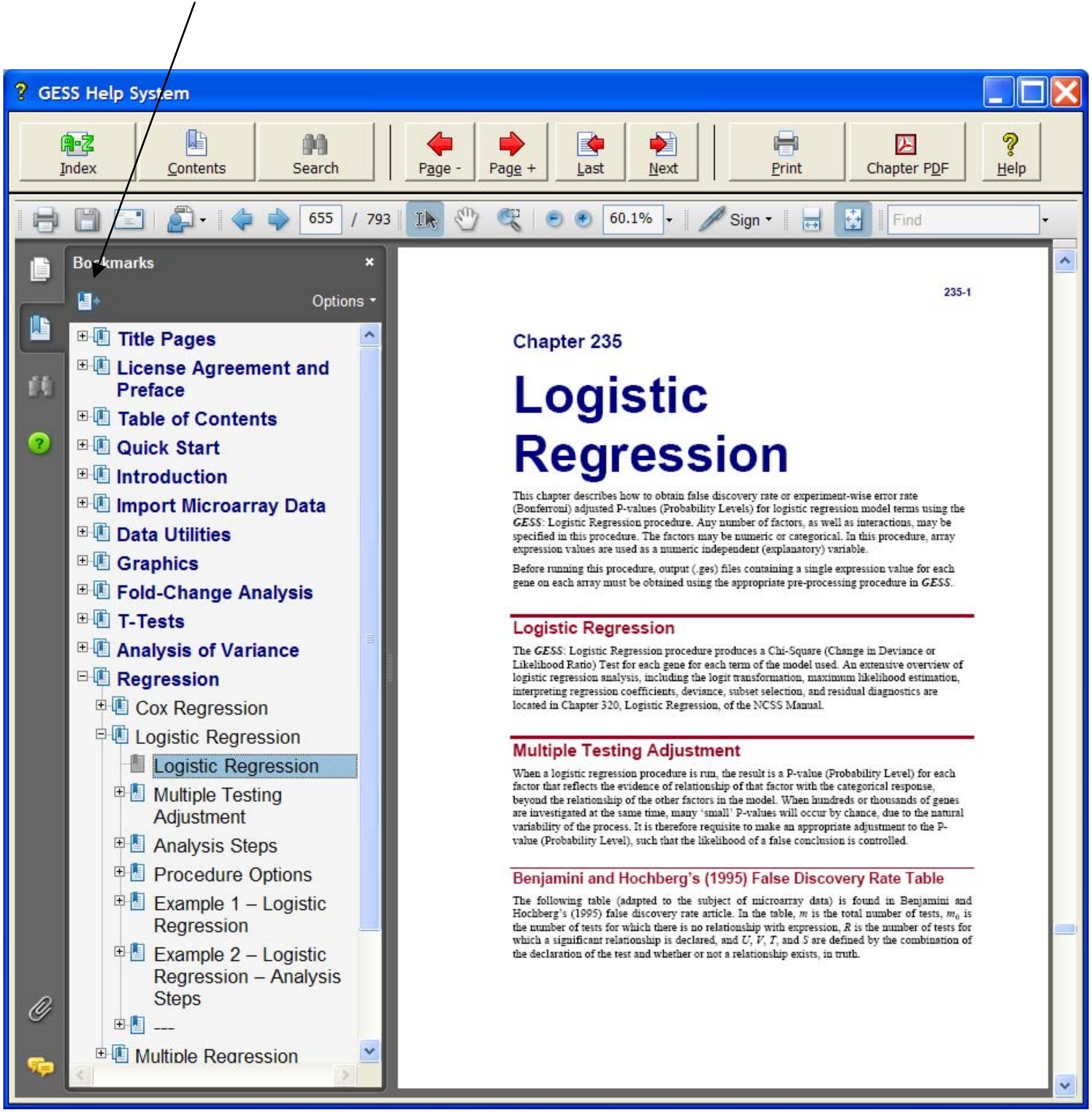
Index entries are displayed in the format

Index Entry --- CHAPTER NAME or CHAPTER NAME --- Index Entry.

You can control which entries are displayed by clicking on the radio buttons at the bottom of the window.

Contents Window

Clicking on the Contents button opens the Contents (Bookmarks) Window of the viewer. From this window you can expand the table of contents to view nested headings. You can click on the "Expand Current Bookmark" icon to instantly find the bookmark location for the currently-displayed page in the help document.



235-1

Chapter 235

Logistic Regression

This chapter describes how to obtain false discovery rate or experiment-wise error rate (Bonferroni) adjusted P-values (Probability Levels) for logistic regression model terms using the *GESS*: Logistic Regression procedure. Any number of factors, as well as interactions, may be specified in this procedure. The factors may be numeric or categorical. In this procedure, array expression values are used as a numeric independent (explanatory) variable.

Before running this procedure, output (.ges) files containing a single expression value for each gene on each array must be obtained using the appropriate pre-processing procedure in *GESS*.

Logistic Regression

The *GESS*: Logistic Regression procedure produces a Chi-Square (Change in Deviance or Likelihood Ratio) Test for each gene for each term of the model used. An extensive overview of logistic regression analysis, including the logit transformation, maximum likelihood estimation, interpreting regression coefficients, deviance, subset selection, and residual diagnostics are located in Chapter 320, Logistic Regression, of the NCSS Manual.

Multiple Testing Adjustment

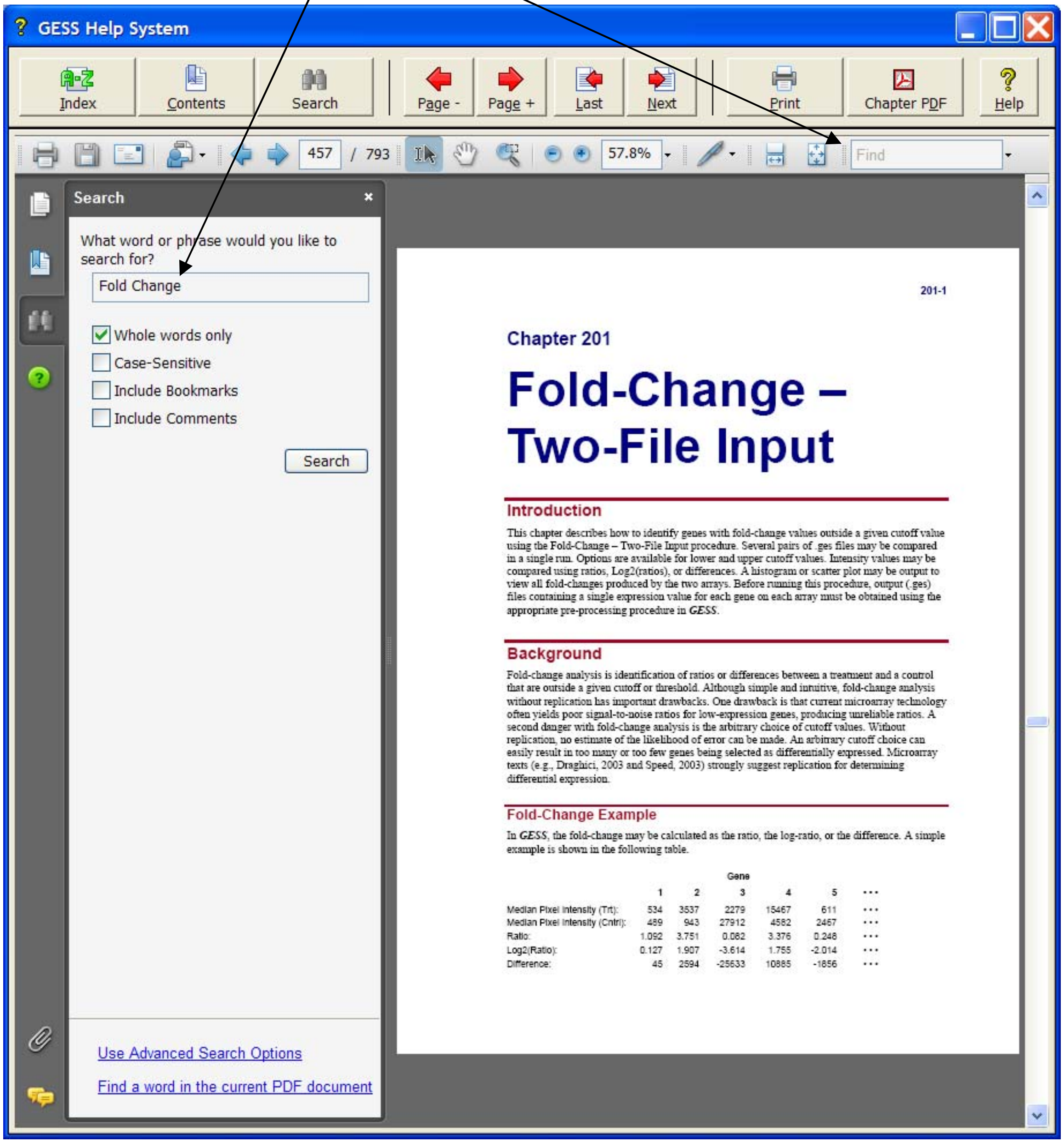
When a logistic regression procedure is run, the result is a P-value (Probability Level) for each factor that reflects the evidence of relationship of that factor with the categorical response, beyond the relationship of the other factors in the model. When hundreds or thousands of genes are investigated at the same time, many 'small' P-values will occur by chance, due to the natural variability of the process. It is therefore requisite to make an appropriate adjustment to the P-value (Probability Level), such that the likelihood of a false conclusion is controlled.

Benjamini and Hochberg's (1995) False Discovery Rate Table

The following table (adapted to the subject of microarray data) is found in Benjamini and Hochberg's (1995) false discovery rate article. In the table, m is the total number of tests, m_0 is the number of tests for which there is no relationship with expression, R is the number of tests for which a significant relationship is declared, and U, V, T, and S are defined by the combination of the declaration of the test and whether or not a relationship exists, in truth.

Search Window

Clicking on the Search button opens the Search Window of the viewer. From this window you can search the entire help system for any word or phrase. A search can also be initiated from the Find box in the viewer toolbar.



What word or phrase would you like to search for?

Fold Change

☑ Whole words only
□ Case-Sensitive
□ Include Bookmarks
□ Include Comments

Search

Use Advanced Search Options

Find a word in the current PDF document

201-1

Chapter 201

Fold-Change – Two-File Input

Introduction

This chapter describes how to identify genes with fold-change values outside a given cutoff value using the Fold-Change – Two-File Input procedure. Several pairs of .ges files may be compared in a single run. Options are available for lower and upper cutoff values. Intensity values may be compared using ratios, Log2(ratios), or differences. A histogram or scatter plot may be output to view all fold-changes produced by the two arrays. Before running this procedure, output (.ges) files containing a single expression value for each gene on each array must be obtained using the appropriate pre-processing procedure in *GESS*.

Background

Fold-change analysis is identification of ratios or differences between a treatment and a control that are outside a given cutoff or threshold. Although simple and intuitive, fold-change analysis without replication has important drawbacks. One drawback is that current microarray technology often yields poor signal-to-noise ratios for low-expression genes, producing unreliable ratios. A second danger with fold-change analysis is the arbitrary choice of cutoff values. Without replication, no estimate of the likelihood of error can be made. An arbitrary cutoff choice can easily result in too many or too few genes being selected as differentially expressed. Microarray texts (e.g., Draghici, 2003 and Speed, 2003) strongly suggest replication for determining differential expression.

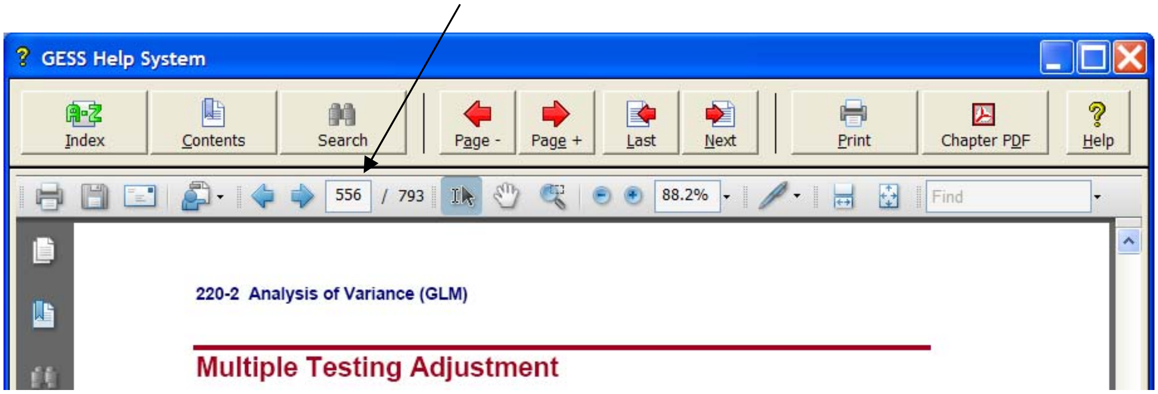
Fold-Change Example

In *GESS*, the fold-change may be calculated as the ratio, the log-ratio, or the difference. A simple example is shown in the following table.

	Gene					
	1	2	3	4	5	• • •
Median Pixel Intensity (Trt):	534	3537	2279	15467	611	• • •
Median Pixel Intensity (Cntrl):	489	943	27912	4582	2467	• • •
Ratio:	1.092	3.751	0.082	3.376	0.248	• • •
Log2(Ratio):	0.127	1.907	-3.614	1.755	-2.014	• • •
Difference:	45	2594	-25633	10885	-1856	• • •

Printing the Documentation

To print pages from the documentation, click on the **Print** button on the ***GESS*** Help System toolbar. This will launch the Adobe Reader print dialogue screen. You can choose to print a single page or a range of pages from the help file. When entering page numbers, remember to use the PDF file page numbers (e.g., 556-558) and not the page numbers found in the document pages (e.g., 220-2 to 220-4 is not a valid page range). The Adobe Reader page numbers can be seen in the viewer window.



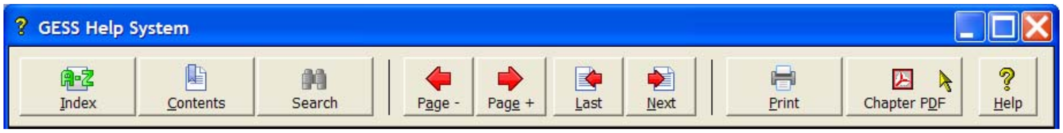
220-2 Analysis of Variance (GLM)

Multiple Testing Adjustment

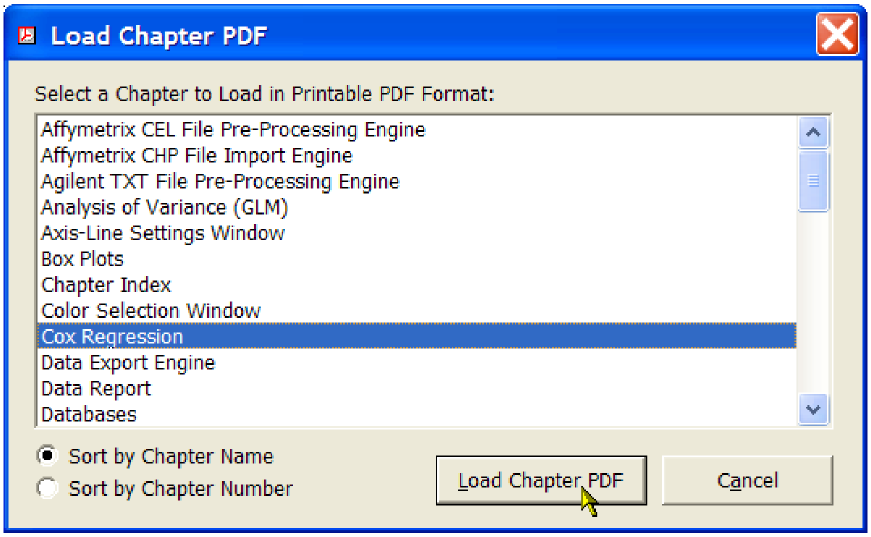
If you are using Adobe Reader 7, then the page numbers are found at the bottom of the viewer window.

To print a single chapter or topic using your default PDF viewer, take the following steps:

1. Click on the **Chapter PDF** icon in the ***NCSS*** Help System toolbar.



2. Choose the chapter you would like to print from the list and click **Load Chapter PDF**. This will launch the individual chapter PDF in a separate window using your default PDF viewer (e.g., Adobe Reader).



Load Chapter PDF

Select a Chapter to Load in Printable PDF Format:

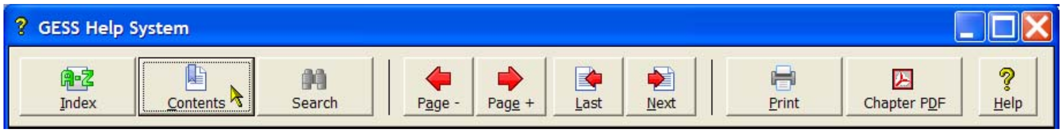
Affymetrix CEL File Pre-Processing Engine
Affymetrix CHP File Import Engine
Agilent TXT File Pre-Processing Engine
Analysis of Variance (GLM)
Axis-Line Settings Window
Box Plots
Chapter Index
Color Selection Window
Cox Regression
Data Export Engine
Data Report
Databases

● Sort by Chapter Name
○ Sort by Chapter Number

3. Use the **Print** function of your PDF viewer to print the entire chapter or individual pages from the chapter.

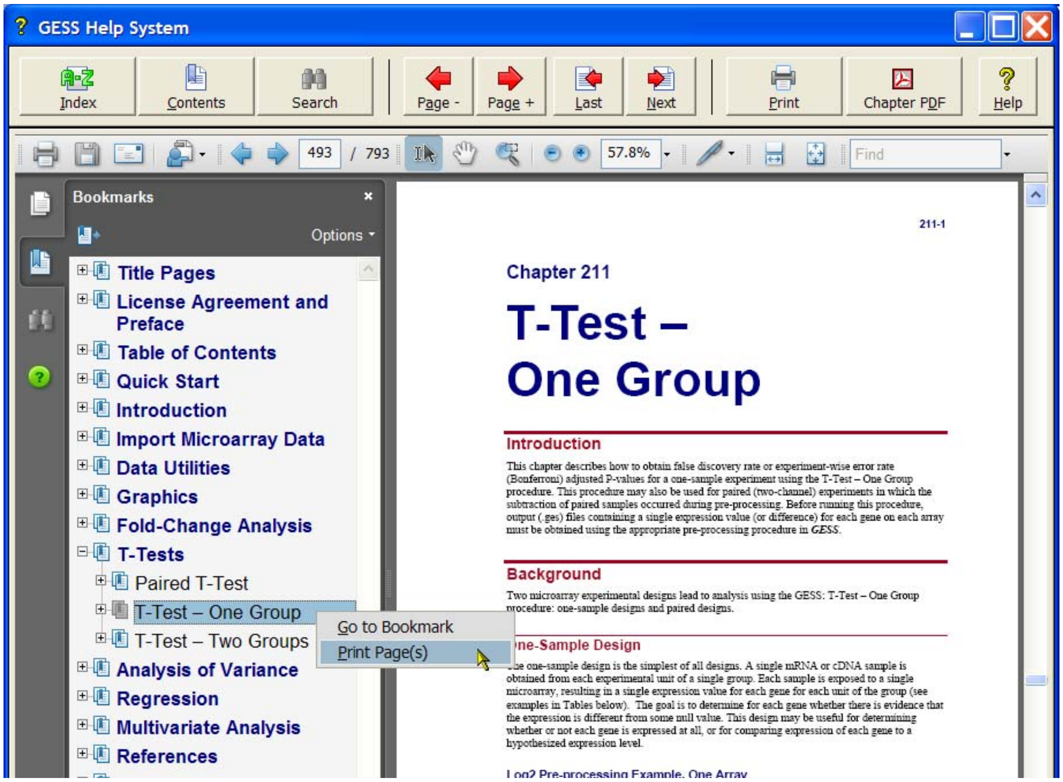
If you have Adobe Reader 8 or later, you can print entire chapters using an alternative method as follows (**This will not work with Adobe Reader 7**):

1. Open the Contents (Bookmarks) Window by clicking on the **Contents button** at the top of the ***GESS*** Help System display window.



2. Expand the bookmarks to display the chapter or topic name you wish to print (e.g., the T-Test – One Group chapter). Then, **highlight** the chapter name, **right-click** on the highlighted selection (or select Options in the panel above), and select **Print Page(s)**. This will automatically print only the pages from the selected chapter.

   CAUTION: When you click Print Page(s), the command is sent to the printer automatically without any intermediate Print Setup window being displayed. Make sure that you have selected only the topic you want before clicking Print Page(s).



211-1

Chapter 211

T-Test –
One Group

Introduction

This chapter describes how to obtain false discovery rate or experiment-wise error rate (Bonferroni) adjusted P-values for a one-sample experiment using the T-Test – One Group procedure. This procedure may also be used for paired (two-channel) experiments in which the subtraction of paired samples occurred during pre-processing. Before running this procedure, output (.ges) files containing a single expression value (or difference) for each gene on each array must be obtained using the appropriate pre-processing procedure in *GESS*.

Background

Two microarray experimental designs lead to analysis using the GESS: T-Test – One Group procedure: one-sample designs and paired designs.

ne-Sample Design

he one-sample design is the simplest of all designs. A single mRNA or cDNA sample is obtained from each experimental unit of a single group. Each sample is exposed to a single microarray, resulting in a single expression value for each gene for each unit of the group (see examples in Tables below). The goal is to determine for each gene whether there is evidence that the expression is different from some null value. This design may be useful for determining whether or not each gene is expressed at all, or for comparing expression of each gene to a hypothesized expression level.

Log2 Pre-processing Example, One Array

If you do not want to print the entire chapter, continue to expand the bookmark tree to the topic you wish to print before completing step 2. The Print Page(s) command prints all pages containing bookmarks that are nested within the highlighted bookmark.

Technical Support

If you have a question about ***GESS***, you should first look to the printed documentation and the included Help system. If you cannot find the answer there, look for help on the web at www.ncss.com/support.html. If you are unable to find the answer to your question by these means, contact ***NCSS*** technical support for assistance by calling (801) 546-0445 between 8 a.m. and 5 p.m. (MST). You can contact us by email at support@ncss.com or by fax at (801) 546-3907. Our technical support staff will help you with your question.

If you encounter problems or errors while using ***GESS***, please view our list of recent corrections before calling by going to www.ncss.com/release_notes.html to find out if you problem or error has been corrected by an update. You can download updates anytime by going to www.ncss.com/download.html. If updating your software does not correct the problem, contact us by phone or email.

To help us answer your questions more accurately, we may need to know about your computer system. Please have pertinent information about your computer and operating system available.

Chapter 2

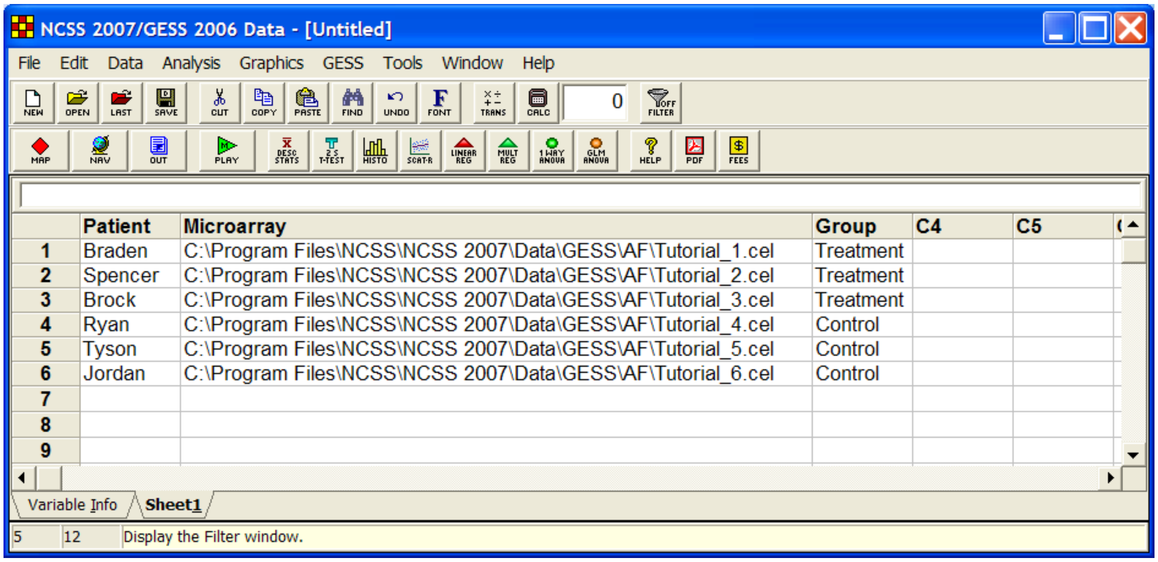
Tutorial

This chapter will quickly familiarize you with the most basic concepts and analysis steps required for a complete microarray analysis using ***GESS***. With the amount of data obtained in microarray experiments, the process of statistical analysis can be very frustrating. ***GESS*** has been designed to take the frustration out of microarray analysis with easy-to-use data entry, importing, pre-processing, and analysis routines. Following a general outline of the analysis steps, we will take you through a step-by-step analysis example, from beginning to end.

Basic Analysis Steps

The following is an outline of the steps involved in a microarray data analysis using ***GESS***. For detailed information about each step, see the tutorial below.

Step 1 – Enter the Microarray Data File Names and Experiment Information into the Spreadsheet



	Patient	Microarray	Group	C4	C5
1	Braden	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel	Treatment		
2	Spencer	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel	Treatment		
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment		
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control		
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control		
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control		
7					
8					
9					

5 12 Display the Filter window.

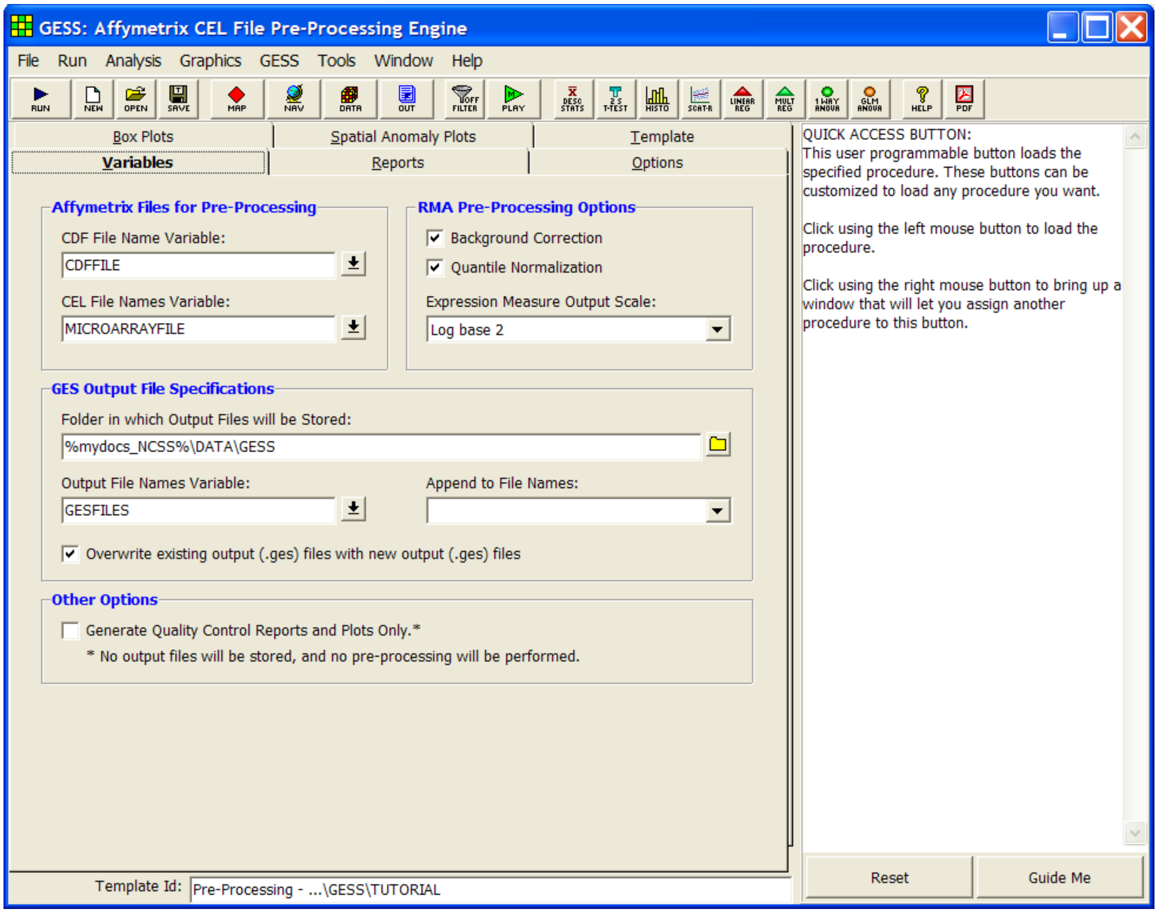
The data system in ***GESS*** is based on a familiar spreadsheet user-interface. Each individual in a microarray experiment is represented by a single row on the database. The microarray data file corresponding to each individual is listed by name in a single column. Data for other variables can be specified as necessary.

The file types that can be imported directly into ***GESS*** are

1. Affymetrix Intensity Files (.cel)
2. Affymetrix Expression Files (.chp)
3. Agilent Intensity Files (.txt)

4. Genepix Intensity Files (.gpr)
5. Generic Two-Channel Intensity Files (.txt)
6. Generic Expression Data Files (.txt, .csv, .dat, .dcp, etc.)

Step 2 – Import the Data Files into GESS and Create .ges Files to Be Used in Further Analyses



Affymetrix Files for Pre-Processing

CDF File Name Variable: CDFFILE

CEL File Names Variable: MICROARRAYFILE

RMA Pre-Processing Options

☑ Background Correction

☑ Quantile Normalization

Expression Measure Output Scale: Log base 2

GES Output File Specifications

Folder in which Output Files will be Stored: %mydocs_NCSS%\DATA\GESS

Output File Names Variable: GESFILES

Append to File Names:

☑ Overwrite existing output (.ges) files with new output (.ges) files

Other Options

□ Generate Quality Control Reports and Plots Only.*

* No output files will be stored, and no pre-processing will be performed.

Template Id: Pre-Processing - ...\GESS\TUTORIAL

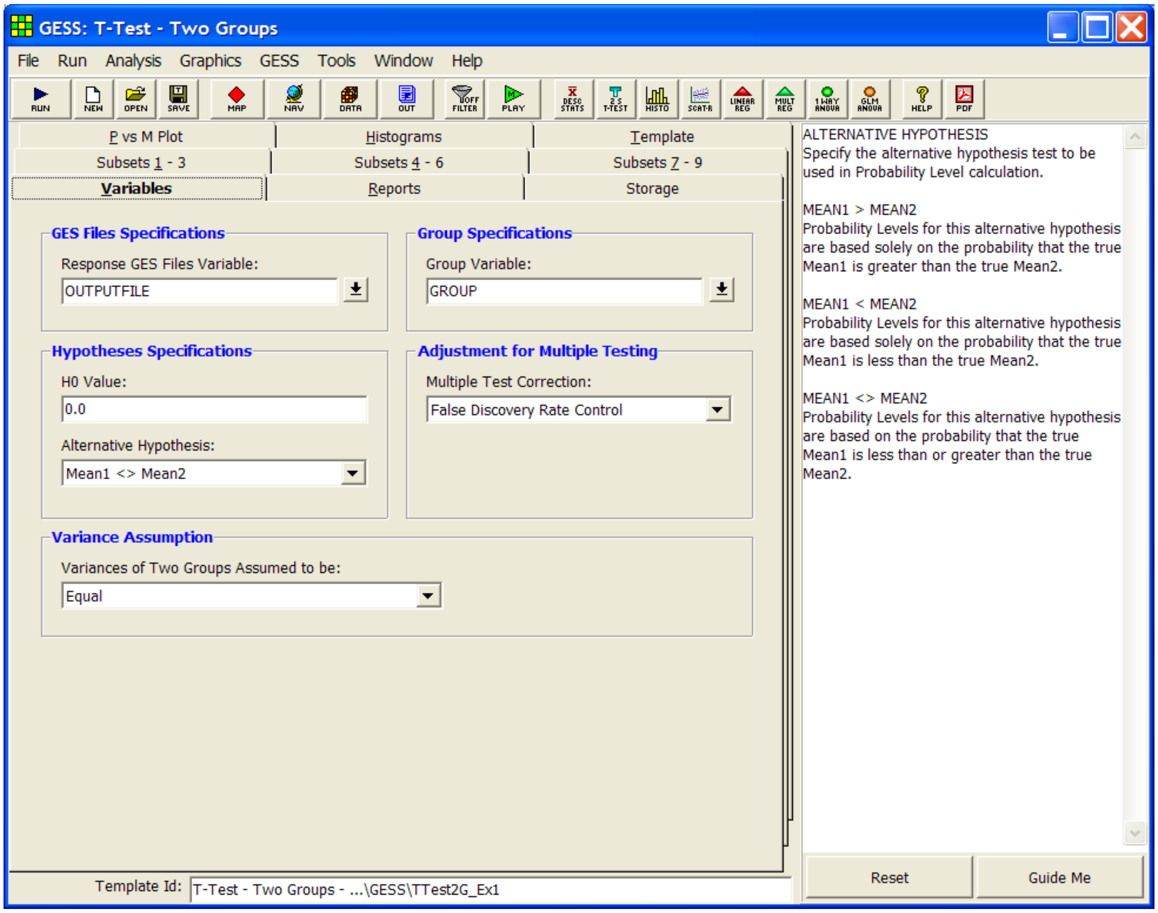
QUICK ACCESS BUTTON:
This user programmable button loads the specified procedure. These buttons can be customized to load any procedure you want.

Click using the left mouse button to load the procedure.

Click using the right mouse button to bring up a window that will let you assign another procedure to this button.

Once the data and file names have been entered into the spreadsheet, you must convert the microarray files (.cel, .chp, .gpr, etc.) into .ges files, the file type used by ***GESS***. The .ges files are read much faster than the original files in subsequent analyses. In the case of Affymetrix .cel files, Agilent .txt files, and GenePix .gpr files, the importing process involves pre-processing as well. A separate import or pre-processing engine is available for each different import-file type.

Step 3 – Perform Statistical Analysis and Output Results



GES Files Specifications

Response GES Files Variable: OUTPUTFILE

Group Specifications

Group Variable: GROUP

Hypotheses Specifications

H0 Value: 0.0

Alternative Hypothesis: Mean1 <> Mean2

Adjustment for Multiple Testing

Multiple Test Correction: False Discovery Rate Control

Variance Assumption

Variances of Two Groups Assumed to be: Equal

Template Id: T-Test - Two Groups - ...\GESS\TTest2G_Ex1

ALTERNATIVE HYPOTHESIS
Specify the alternative hypothesis test to be used in Probability Level calculation.

MEAN1 > MEAN2
Probability Levels for this alternative hypothesis are based solely on the probability that the true Mean1 is greater than the true Mean2.

MEAN1 < MEAN2
Probability Levels for this alternative hypothesis are based solely on the probability that the true Mean1 is less than the true Mean2.

MEAN1 <> MEAN2
Probability Levels for this alternative hypothesis are based on the probability that the true Mean1 is less than or greater than the true Mean2.

After importing and pre-processing, you can proceed to the completion of statistical analysis. The statistical routines in ***GESS*** automatically adjust the p-values for multiple testing (if desired). To complete the analysis, simply fill out the procedure window and click Run. The output listing the significant genes and displaying powerful graphics will automatically be displayed. You can then save the data from significant genes to the spreadsheet for further analysis using ***NCSS***, or cut and paste the report directly into a document or presentation. You also have the option to specify subsets of genes for analysis, thus, allowing you to narrow the list of target genes and increase your power for detecting significant differences or effects.

Several statistical routines are available in ***GESS***, each with full documentation:

1. Fold-Change Analysis
2. Paired, One-Sample and Two-Sample T-Tests
3. Analysis of Variance (GLM)
4. Repeated Measures ANOVA
5. Cox Regression
6. Logistic Regression
7. Multiple Regression
8. Principal Components Analysis
9. Hierarchical Cluster Analysis

Tutorial

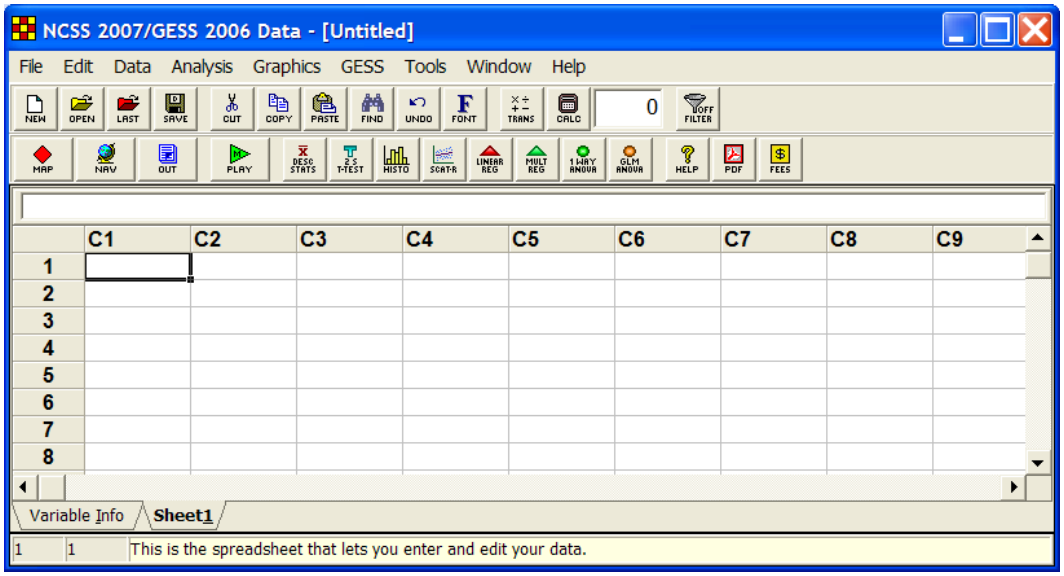
We will now take you through the analysis process step by step. In this analysis we will use a sample Affymetrix dataset consisting of six .cel files. Our goal will be to perform a T-test for differential expression and a follow-up cluster analysis on significant genes. This tutorial will show you how a general statistical analysis is done, from a blank spreadsheet to a final report. While the individual steps may vary for other microarray platforms, e.g. GenePix, Agilent, etc., the overall analysis process is the same. Specific information relating to each importing and analysis procedure can be found in other chapters of this manual.

Step 1 – Enter the Microarray Data File Names and Experiment Information into the Spreadsheet

We will first present the basic steps necessary to start ***GESS***, change the column names, and enter data and file names. The input files for this tutorial are stored in the *C:\Program Files\NCSS \NCSS 2007\Data\GESS\AF* directory. After completing this tutorial, your dataset should match the TUTORIAL dataset found in the *C:\…\[My] Documents\NCSS\NCSS 2007\Data\GESS* directory.

1 Launch GESS.

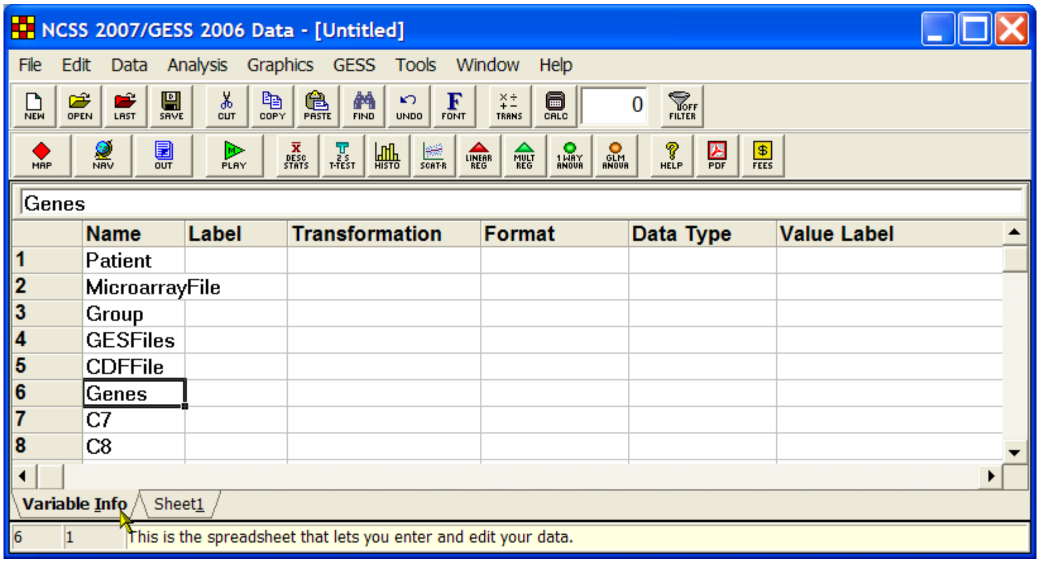
- Double click the ***GESS*** desktop icon or launch ***GESS*** from the Windows Start menu by clicking on **All Programs**. The GESS Data window will appear.



This is the spreadsheet that lets you enter and edit your data.

2 Change the Column Names.

- Click on the **Variable Info** tab in the bottom left-hand corner of the screen.
- In the **Name** column, select the cell containing **C1**, and enter **Patient**. Enter **MicroarrayFile**, **Group**, **GESFiles**, **CDFFile**, and **Genes** for **C2-C6**, respectively. We will fill in these variables as we go through this tutorial.
- Click on the **Sheet1 tab** at the bottom to return to the data sheet. The columns now have the new names as their titles.


NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 FILTER
MAP NAV OUT PLAY DESC STATS 2 S T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info Sheet1

Genes

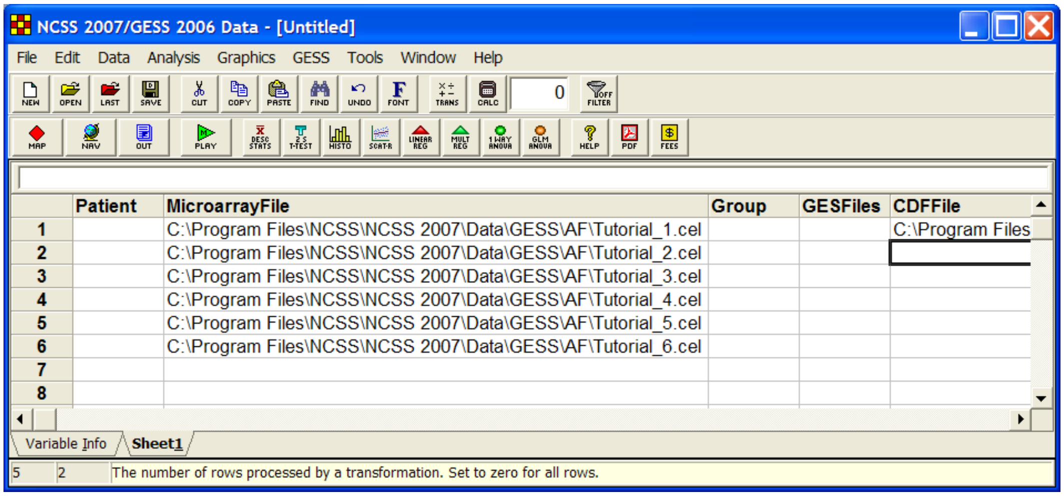
	Name	Label	Transformation	Format	Data Type	Value Label
1	Patient					
2	MicroarrayFile					
3	Group					
4	GESFiles					
5	CDFFile					
6	Genes					
7	C7					
8	C8					

6 1 This is the spreadsheet that lets you enter and edit your data.

3 Enter the Names of Existing Microarray Files for Analysis.

- Highlight the first cell in the **MicroarrayFile** column. Select **Edit**, then **Enter File Name**, or hit **F7** to launch the file browser. Under **Save as type** select **Affymetrix (*.CEL)**.
- Browse to the folder into which you installed ***GESS*** (usually *C:\Program Files\NCSS \NCSS 2007*).
- Select the **DATA** subdirectory of the **NCSS 2007** directory.
- Open the **GESS** folder and then the **AF** folder.
- Select the file **Tutorial_1.cel** and click **Save**. This will save the file name and path to the spreadsheet. Repeat the preceding steps for **Tutorial_2.cel through Tutorial_6.cel**. Hint: Using **F7** to launch the file browser makes this process go very fast.
- Highlight the first cell in the **CDFFile** column. Select **Edit**, then **Enter File Name**, or hit **F7** to launch the file browser. Under **Save as type** select **Affymetrix (*.CDF)**.
- Browse to the folder into which you installed ***GESS*** (usually *C:\Program Files\NCSS \NCSS 2007*).
- Select the **DATA** subdirectory of the **NCSS 2007** directory.
- Open the **GESS** folder and then the **AF** folder.
- Select the file **Test3.cdf** and click **Save**. This will save the file name and path to the spreadsheet.

  Note: The .cdf file is only required when importing Affymetrix .cel or .chp files, and must be obtained independently of the ***GESS*** system before you can perform your analysis. You can download Affymetrix .cdf library files from www.Affymetrix.com.
- Expand the Microarray_File and CDFFile column widths by choosing **Edit**, the Re**size Rows and Columns**, then **Resize using Data and Titles** from the Data window menus. This will allow you to see the entire file path when more data are entered.

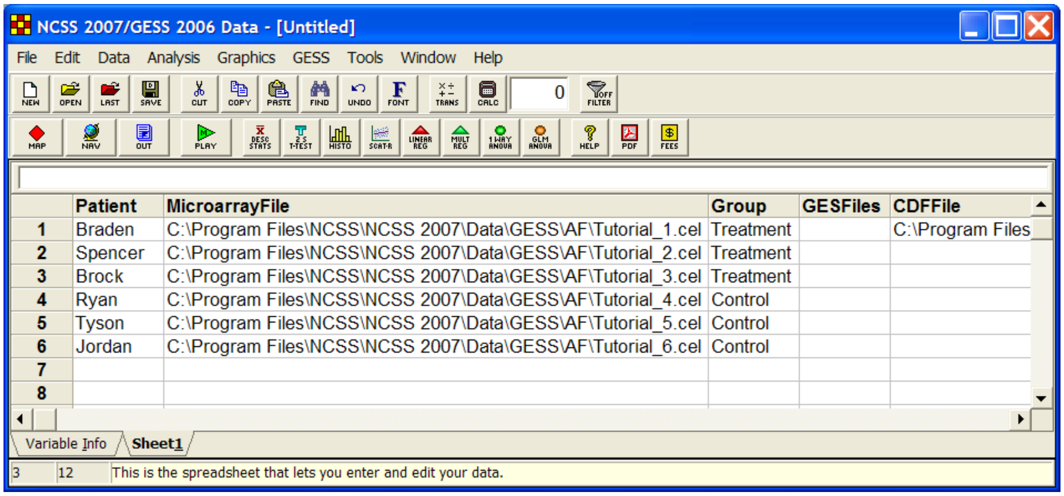

NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 FILTER
MAP NAV OUT PLAY DESC STATS T-TEST HISTO SCAT-R LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info
Sheet1

	Patient	MicroarrayFile	Group	GESFiles	CDFFile
1		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel			C:\Program Files
2		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel			
3		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel			
4		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel			
5		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel			
6		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel			
7					
8					

5 2 The number of rows processed by a transformation. Set to zero for all rows.

4 Enter the Experiment Data.

- In the **Patient** column, enter **Braden**, **Spencer**, **Brock**, **Ryan**, **Tyson**, and **Jordan**, as the individual names.
- In the **Group** column, enter **Treatment** for the first three rows and **Control** for the last three rows.
- Save the data file as **TUTORIAL.S0** by clicking on **File** and then **Save**. The data are now ready for pre-processing and analysis.


NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 FILTER
MAP NAV OUT PLAY DESC STATS T-TEST HISTO SCAT-R LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info
Sheet1

	Patient	MicroarrayFile	Group	GESFiles	CDFFile
1	Braden	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel	Treatment		C:\Program Files
2	Spencer	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel	Treatment		
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment		
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control		
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control		
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control		
7					
8					

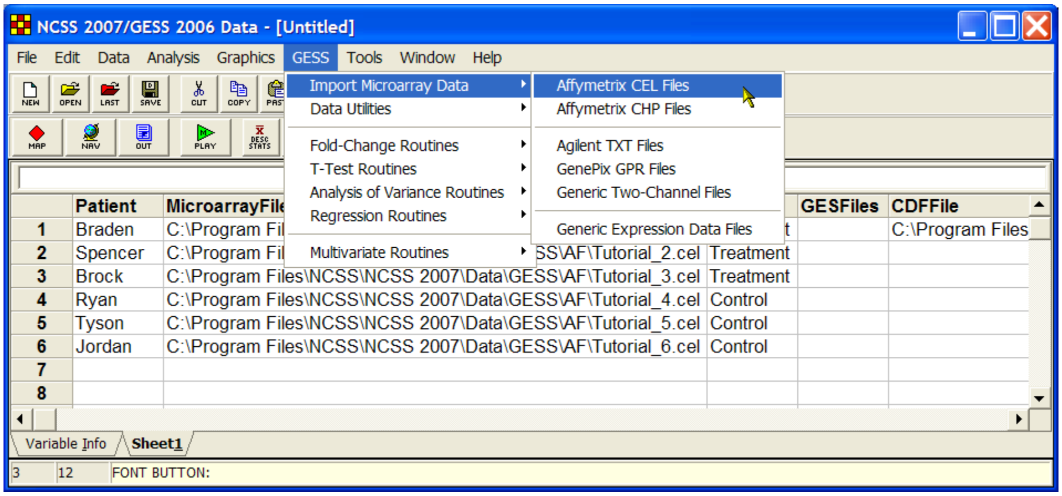
3 12 This is the spreadsheet that lets you enter and edit your data.

Step 2 – Import the Data Files into GESS and Create .ges Files to Be Used in Further Analyses

We now present the steps necessary to import files into ***GESS***. The input files for this tutorial are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF* directory. To run this example, take the following steps or load the **GESS Tutorial - Step 2** template on the Affymetrix CEL File Pre-Processing Engine Template tab.

1 Launch the Importing or Pre-Processing Procedure.

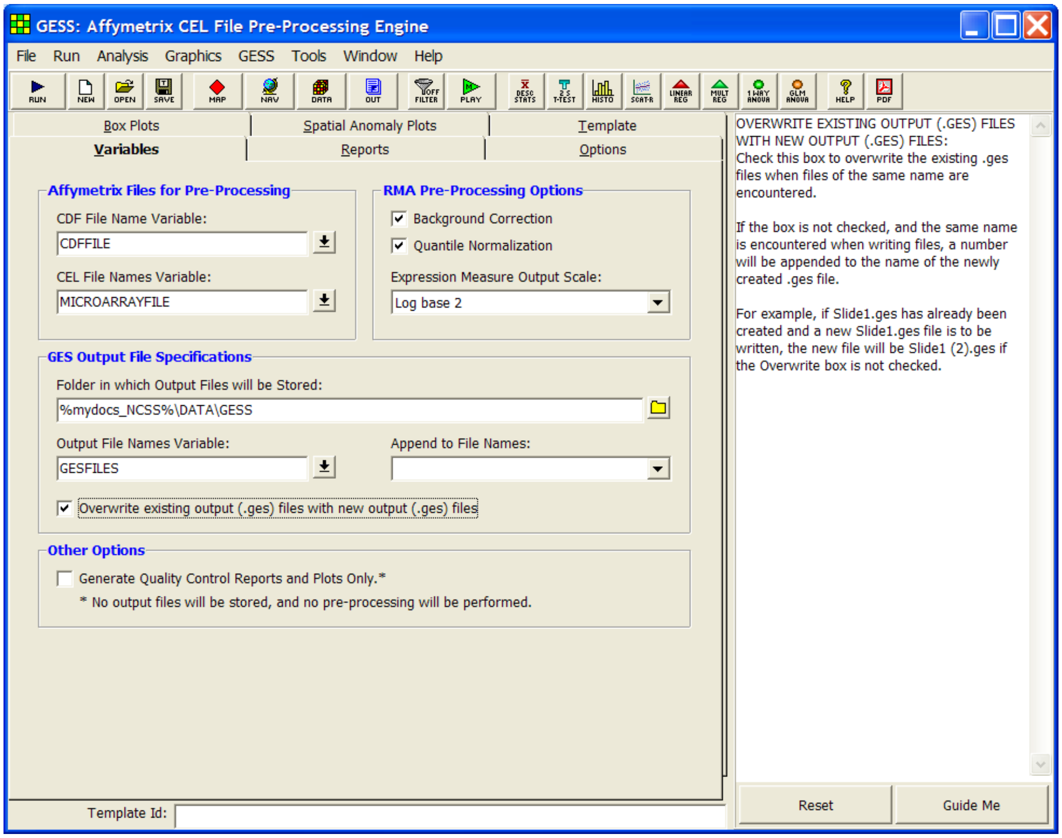
- On the menus, select **GESS**, then **Import Microarray Data**, then **Affymetrix CEL Files**. The Affymetrix CEL File Pre-Processing Engine procedure will be displayed. Hint: Right click on one of the eight quick-launch icons on the spreadsheet toolbar to select icons that you use most frequently. The new icons will be placed on the toolbar.
- On the Affymetrix CEL File Pre-Processing Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.



	Patient	MicroarrayFile		GESFiles	CDFFile
1	Braden	C:\Program Fi	t		C:\Program Files
2	Spencer	C:\Program Fil … SS\AF\Tutorial_2.cel	Treatment		
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment		
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control		
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control		
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control		
7					
8					

2 Specify the Variables.

- On the Affymetrix CEL File Pre-Processing Engine window, select the **Variables tab**.
- Set the **CDF File Name Variable** to **CDFFile**.
- Set the **CEL File Names Variable** to **MicroarrayFile**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**. The designator "%mydocs_NCSS%" represents the path to the *NCSS* personal data folder (commonly *C:\...\[My] Documents\NCSS\NCSS 2007*).
- Set the **Output File Names Variable** to **GESFiles**.
- Check the box next to **Overwrite existing output (.ges) files with new output (.ges) files**.
- Leave all other options under the Variables tab at their default settings.



Affymetrix Files for Pre-Processing

CDF File Name Variable: CDFFILE

CEL File Names Variable: MICROARRAYFILE

RMA Pre-Processing Options

☑ Background Correction

☑ Quantile Normalization

Expression Measure Output Scale: Log base 2

GES Output File Specifications

Folder in which Output Files will be Stored: %mydocs_NCSS%\DATA\GESS

Output File Names Variable: GESFILES

Append to File Names:

☑ Overwrite existing output (.ges) files with new output (.ges) files

Other Options

☐ Generate Quality Control Reports and Plots Only.*

* No output files will be stored, and no pre-processing will be performed.

Template Id:

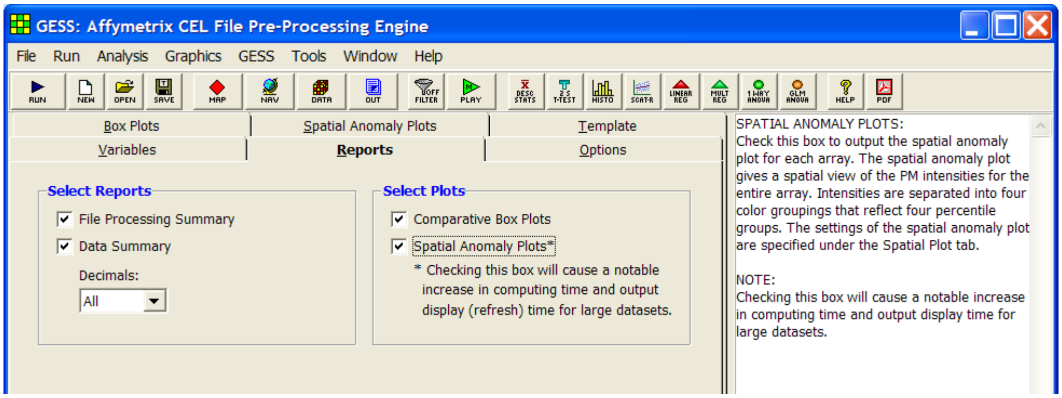
OVERWRITE EXISTING OUTPUT (.GES) FILES WITH NEW OUTPUT (.GES) FILES:
Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file.

For example, if Slide1.ges has already been created and a new Slide1.ges file is to be written, the new file will be Slide1 (2).ges if the Overwrite box is not checked.

3 Specify the Reports.

- Select the **Reports tab**.
- Check the box next to **Spatial Anomaly Plots**.
- Leave all other options under the Reports tab and other tabs at their default settings.



Select Reports

☑ File Processing Summary

☑ Data Summary

Decimals: All

Select Plots

☑ Comparative Box Plots

☑ Spatial Anomaly Plots*

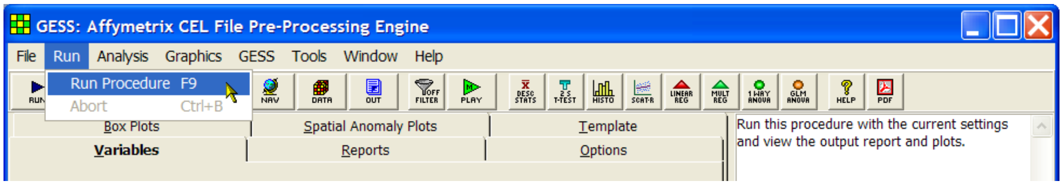
* Checking this box will cause a notable increase in computing time and output display (refresh) time for large datasets.

SPATIAL ANOMALY PLOTS:
Check this box to output the spatial anomaly plot for each array. The spatial anomaly plot gives a spatial view of the PM intensities for the entire array. Intensities are separated into four color groupings that reflect four percentile groups. The settings of the spatial anomaly plot are specified under the Spatial Plot tab.

NOTE:
Checking this box will cause a notable increase in computing time and output display time for large datasets.

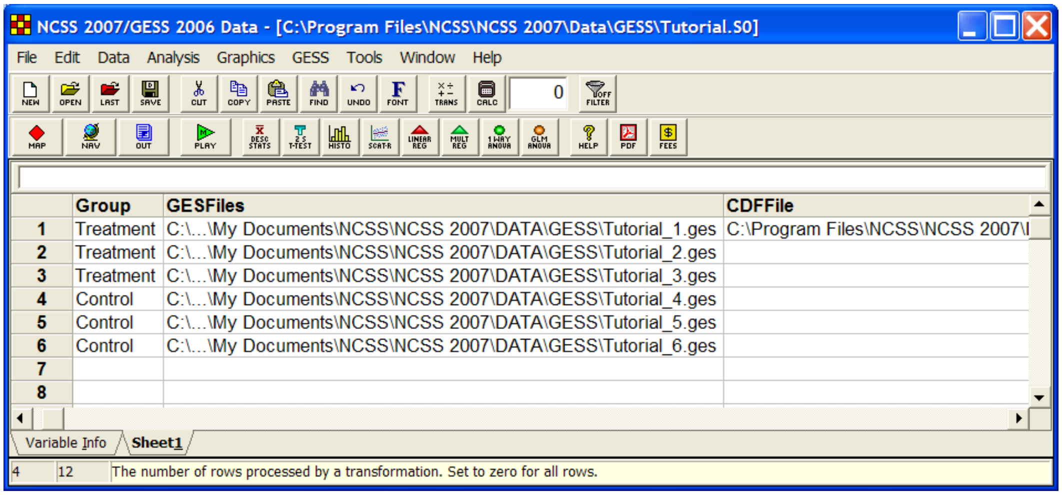
4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the toolbar at the top).



Run this procedure with the current settings and view the output report and plots.

- The **GESFiles** column on the spreadsheet now appears as follows (after resizing the column width):



	Group	GESFiles	CDFFile
1	Treatment	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_1.ges	C:\Program Files\NCSS\NCSS 2007\I
2	Treatment	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_2.ges	
3	Treatment	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_3.ges	
4	Control	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_4.ges	
5	Control	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_5.ges	
6	Control	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_6.ges	
7			
8			

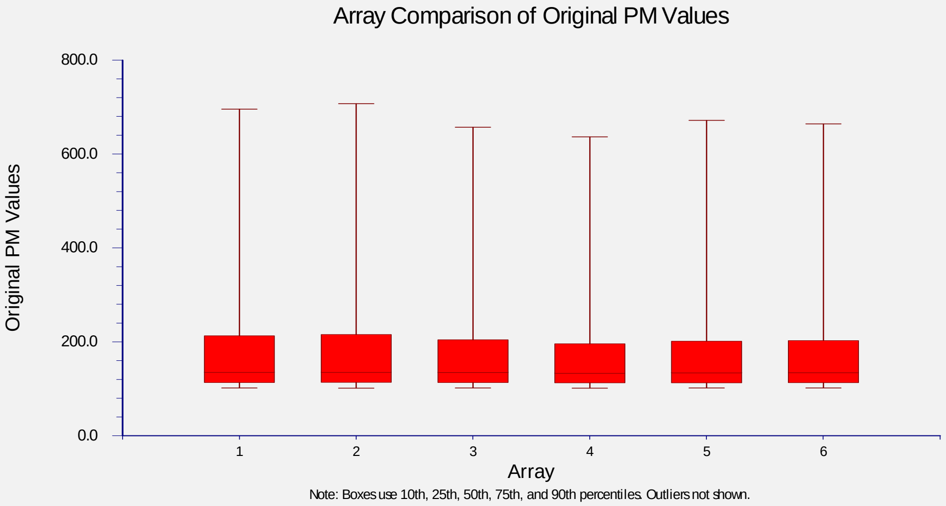
4 12 The number of rows processed by a transformation. Set to zero for all rows.

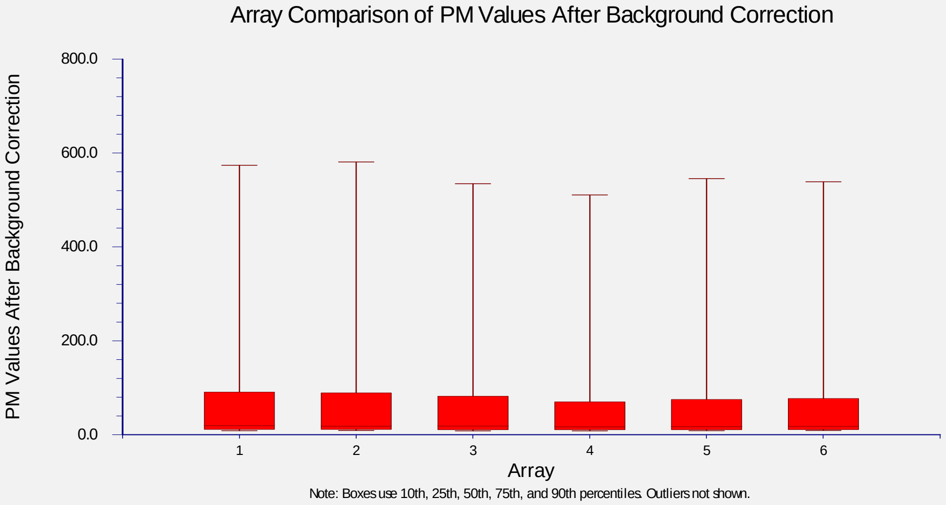
Notice that the **GESFiles** column has been automatically filled with the names of the newly created .ges files. These .ges files will be used by ***GESS*** when array data is needed in future analyses. The report output is given below.

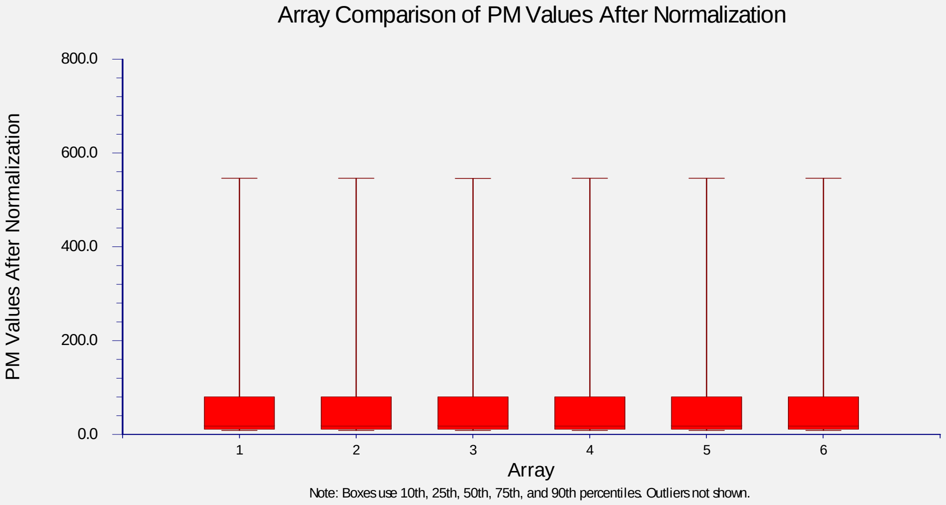
Report Output

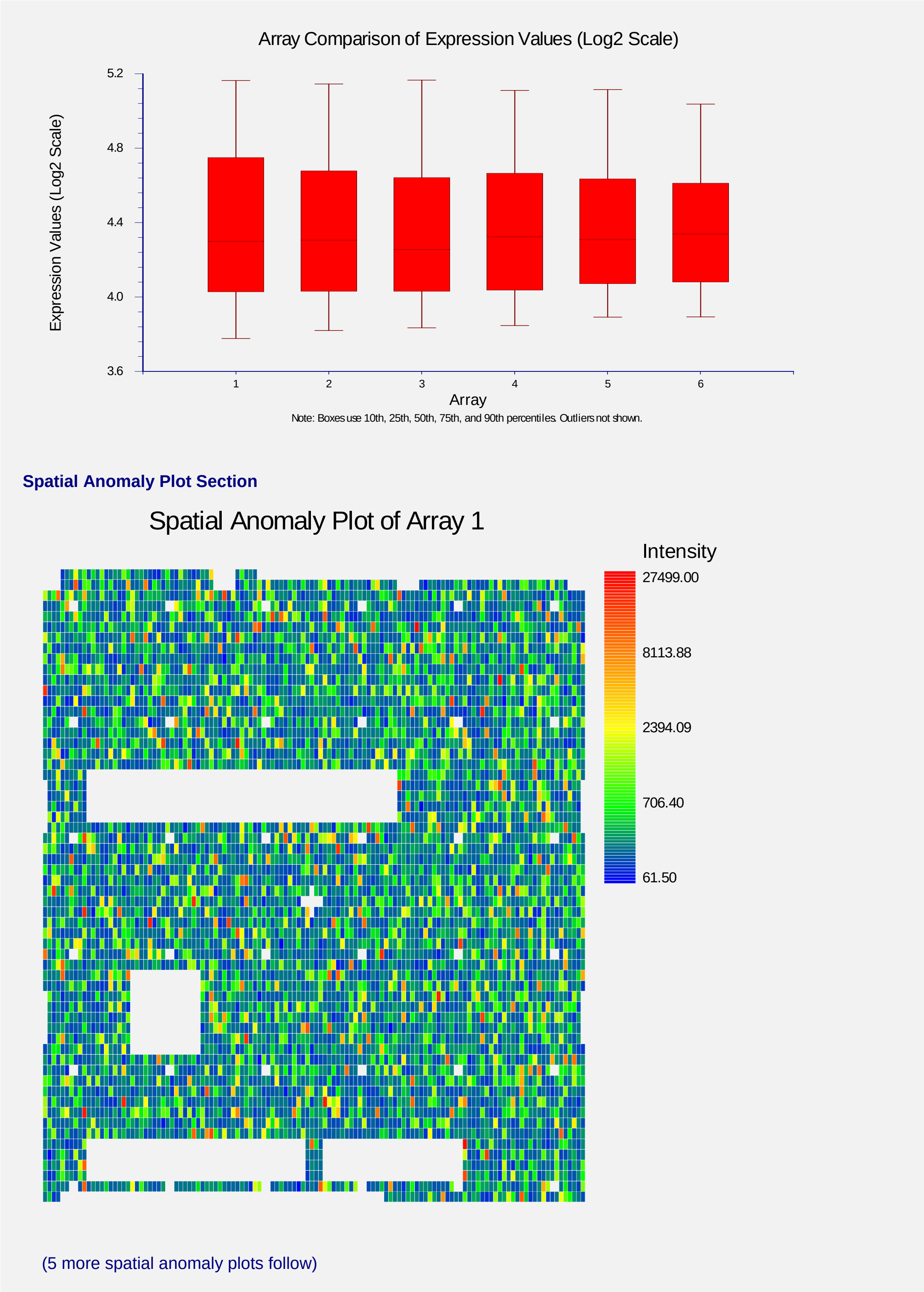
```
Input File Summary

             Number of
Row          Probes       File Name
1            15876        C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel
2            15876        C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel
3            15876        C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel
4            15876        C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel
5            15876        C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel
6            15876        C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel

CDF File Name: C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Test3.cdf
CDH File Name: C:\...\My Documents\NCSS\NCSS 2007\Data\GESS\CDH\Test3.cdh


Pre-Processing Methods Summary

Task Name                  Method Selected
Background Correction      RMA (Model-Based)
Normalization              Quantile
Summarization              Median Polish
Output Scale               Log base 2


Output File Summary

             Number of
Row          Probe Sets   File Name
1            345          C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_1.ges
2            345          C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_2.ges
3            345          C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_3.ges
4            345          C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_4.ges
5            345          C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_5.ges
6            345          C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_6.ges


Numerical Summary of Original PM Intensities

                          10th        25th        50th        75th        90th
Row   Minimum       Percentile  Percentile  Percentile  Percentile  Percentile     Maximum
1        61.5          101.8       113.5         136       212.8      695.65       27499
2        64.3          101.3       113.8       135.8       215.5         707     27151.3
3        70.5         101.65       113.3       135.5       204.3       656.8     23705.8
4        63.3          101.5       112.8       133.8       195.8      636.55     37486.5
5        65.3          101.8       112.8       134.8       201.3      671.65     31738.5
6          66         101.65         113       135.3       202.8       664.3       44675


Numerical Summary of Expression Values (Log2 Scale)

                          10th        25th        50th        75th        90th
Row   Minimum       Percentile  Percentile  Percentile  Percentile  Percentile     Maximum
1     3.432443       3.777095    4.028214     4.30166    4.749565    5.163875     9.09646
2      3.37914       3.819739    4.030816    4.307514    4.677792    5.144782    9.067124
3     3.472117       3.834185    4.031287    4.257486    4.641731    5.165857    9.163662
4     3.358151       3.846722    4.037156    4.325556    4.664896    5.110379    9.000655
5     3.435157       3.892049    4.071929    4.310611    4.635222    5.114842    9.055325
6     3.243854       3.893716     4.08064    4.341212    4.611947    5.037131    8.962646
```

Box Plot Section









Spatial Anomaly Plot Section

(5 more spatial anomaly plots follow)

The reports give you a summary of the files processed and the data they contain, as well as graphical representations of the individual array data. Double-click on any plot to enlarge it.

Step 3 – Perform Statistical Analyses and Output Results

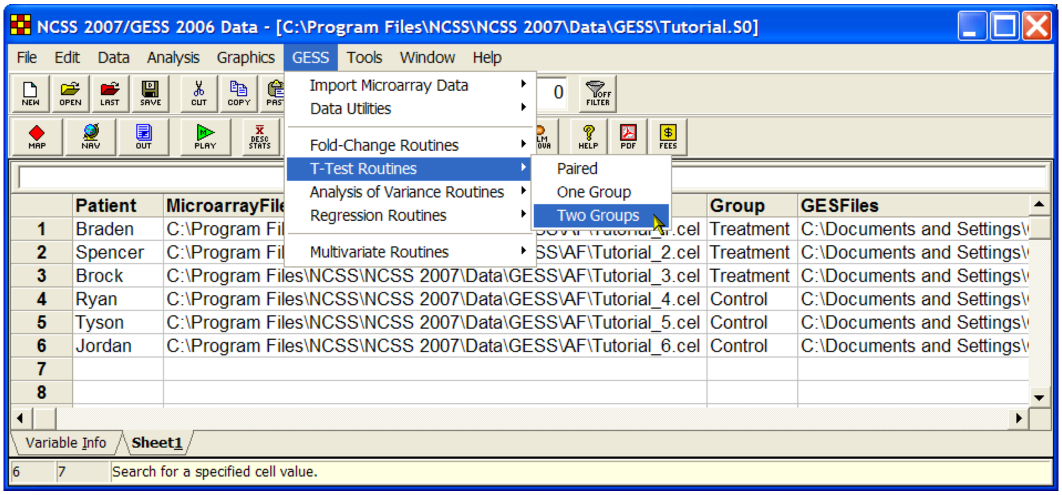
We now present some basic statistical analyses in ***GESS***. We will analyze the data using a two-sample T-test, followed by a hierarchical cluster analysis of significant genes.

Two-Sample T-Test Steps

To run this example, take the following steps or load the **GESS Tutorial - Step 3** template on the T-Test - Two Groups Template tab.

1 **Launch the T-Test - Two Groups Procedure.**

- On the menus, select **GESS**, then **T-Test Routines**, then **Two Groups**. The T-Test - Two Groups procedure will be displayed. Hint: Right click on one of the eight quick-launch icons on the spreadsheet toolbar to select icons that you use most frequently. The new icons will be placed on the toolbar.
- On the T-Test - Two Groups window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

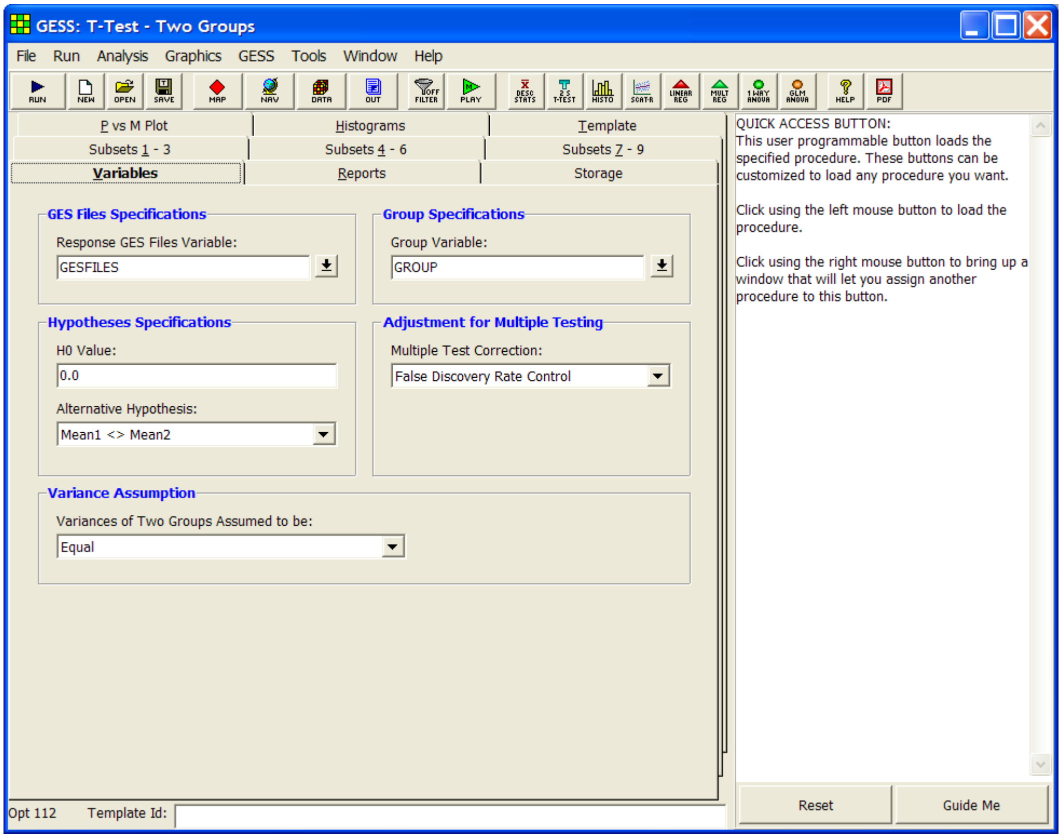


	Patient	MicroarrayFile	Group	GESFiles
1	Braden	C:\Program Fil ... .cel	Treatment	C:\Documents and Settings\
2	Spencer	C:\Program Fil ... SS\AF\Tutorial_2.cel	Treatment	C:\Documents and Settings\
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment	C:\Documents and Settings\
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control	C:\Documents and Settings\
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control	C:\Documents and Settings\
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control	C:\Documents and Settings\
7				
8				

6 7 Search for a specified cell value.

2 **Specify the Variables.**

- On the T-Test - Two Groups window, select the **Variables tab**.
- Set the **Response GES Files Variable** to **GESFiles**.
- Set the **Group Variable** to **Group**.
- Leave all other options under the Variables tab at their default settings.


GESS: T-Test - Two Groups
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS 2 S T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
P vs M Plot Histograms Template
Subsets 1 - 3 Subsets 4 - 6 Subsets 7 - 9
Variables Reports Storage
Reset
Guide Me
Opt 112 Template Id:

GES Files Specifications

Response GES Files Variable: GESFILES

Group Specifications

Group Variable: GROUP

Hypotheses Specifications

H0 Value: 0.0

Alternative Hypothesis: Mean1 <> Mean2

Adjustment for Multiple Testing

Multiple Test Correction: False Discovery Rate Control

Variance Assumption

Variances of Two Groups Assumed to be: Equal

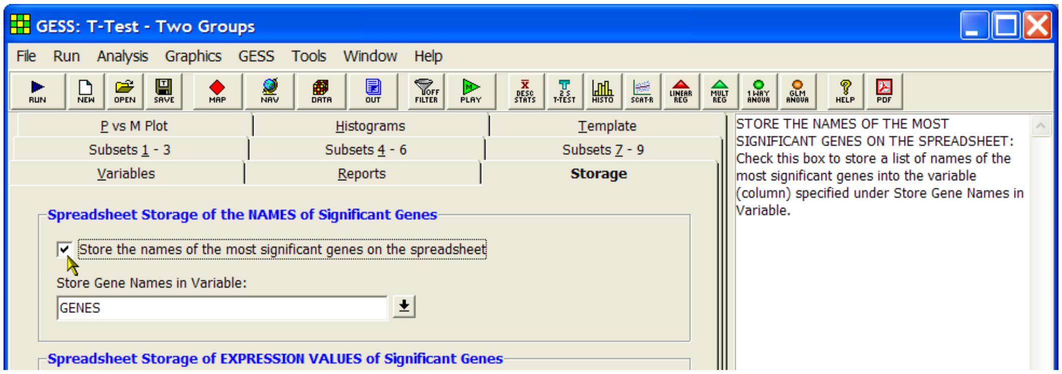
QUICK ACCESS BUTTON:
This user programmable button loads the specified procedure. These buttons can be customized to load any procedure you want.

Click using the left mouse button to load the procedure.

Click using the right mouse button to bring up a window that will let you assign another procedure to this button.

3 Specify the Storage Data.

- Select the **Storage tab**.
- Check the box next to **Store the names of the most significant genes on the spreadsheet**.
- Set the **Store Gene Names in Variable** to **Genes**.
- Leave all other options under the Storage tab and other tabs at their default settings.


GESS: T-Test - Two Groups
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS 2 S T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
P vs M Plot Histograms Template
Subsets 1 - 3 Subsets 4 - 6 Subsets 7 - 9
Variables Reports Storage

Spreadsheet Storage of the NAMES of Significant Genes

☑ Store the names of the most significant genes on the spreadsheet

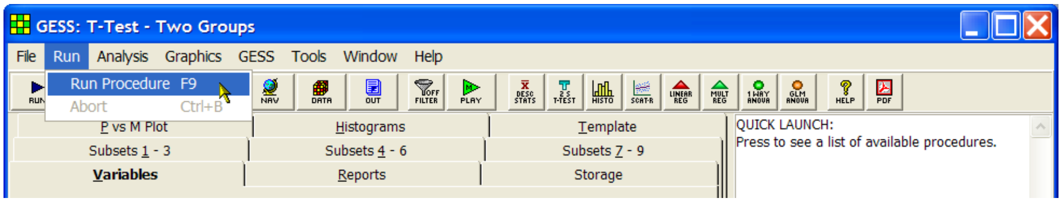
Store Gene Names in Variable: GENES

Spreadsheet Storage of EXPRESSION VALUES of Significant Genes

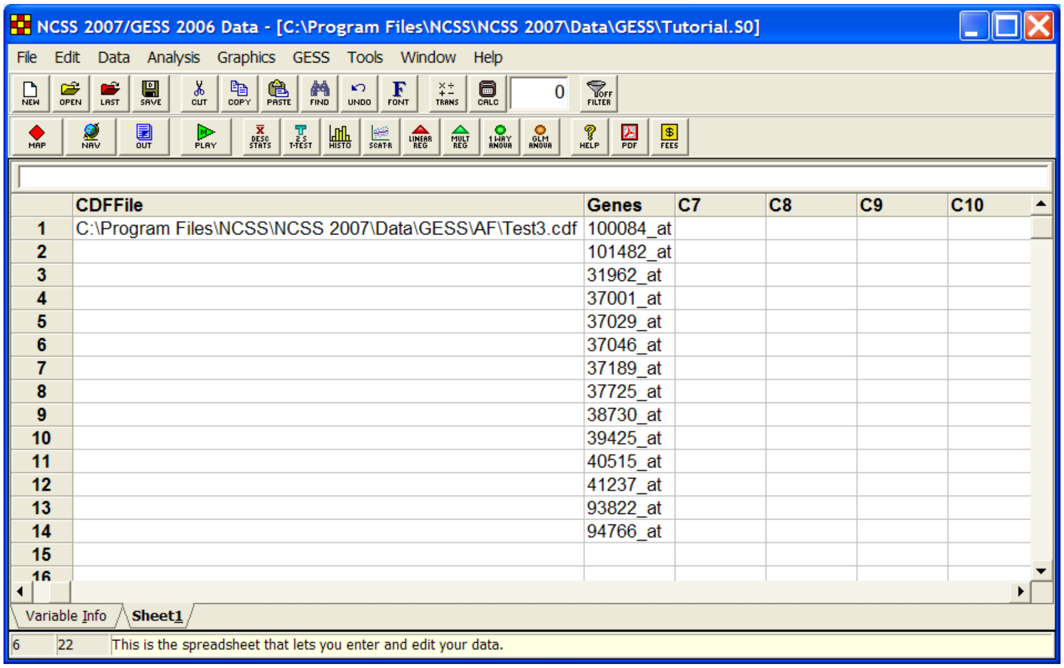
STORE THE NAMES OF THE MOST SIGNIFICANT GENES ON THE SPREADSHEET: Check this box to store a list of names of the most significant genes into the variable (column) specified under Store Gene Names in Variable.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the toolbar at the top).


GESS: T-Test - Two Groups
File Run Analysis Graphics GESS Tools Window Help
RUN
Run Procedure F9
Abort Ctrl+B
NAV DATA OUT FILTER PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
P vs M Plot
Histograms
Template
Subsets 1 - 3
Subsets 4 - 6
Subsets 7 - 9
Variables
Reports
Storage
QUICK LAUNCH:
Press to see a list of available procedures.

The **Genes** column on the spreadsheet now appears as follows (after resizing the column width):

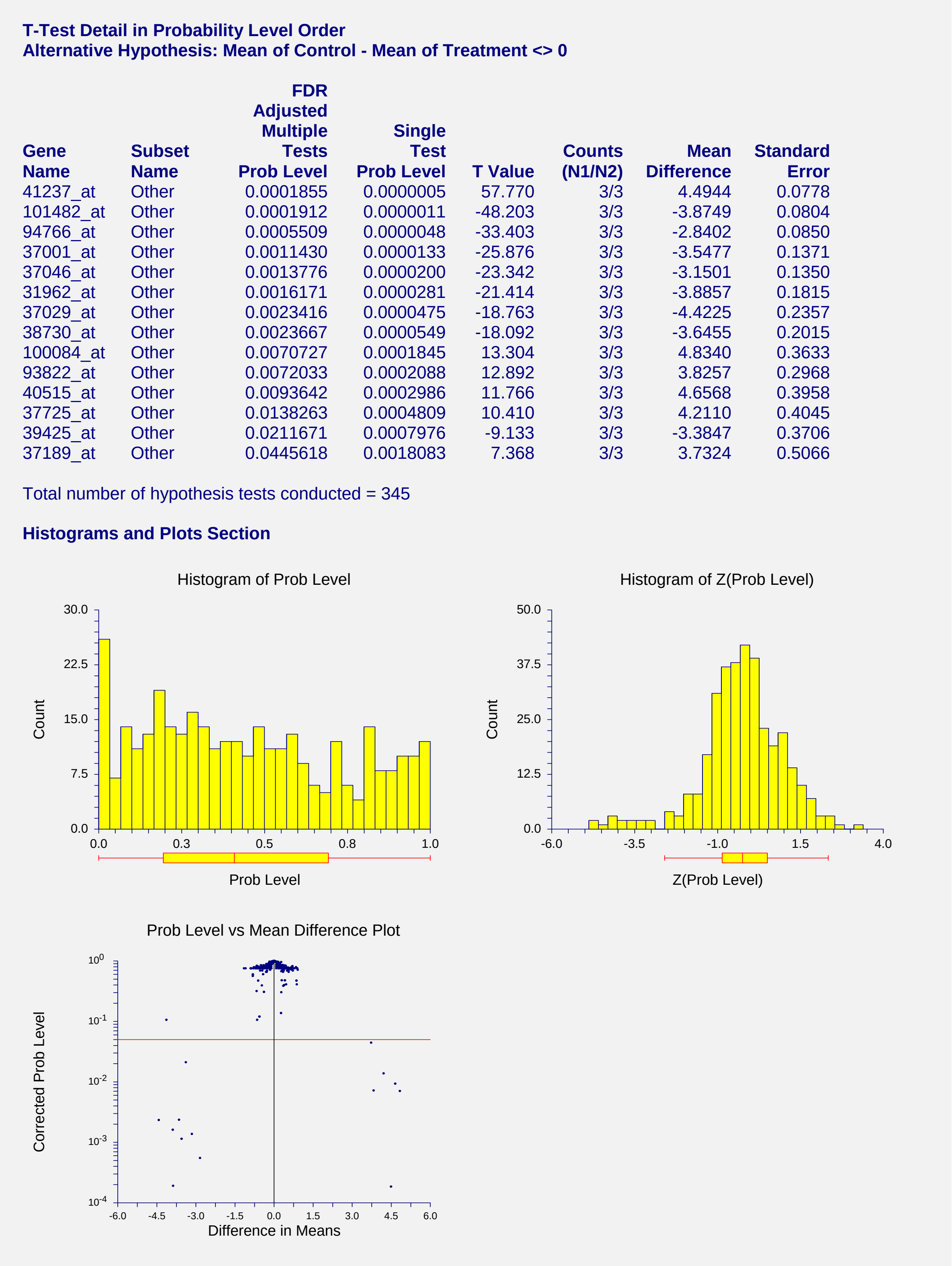

NCSS 2007/GESS 2006 Data - [C:\Program Files\NCSS\NCSS 2007\Data\GESS\Tutorial.S0]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 FILTER
MAP NAV OUT PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info
Sheet1

	CDFFile	Genes	C7	C8	C9	C10
1	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Test3.cdf	100084_at				
2		101482_at				
3		31962_at				
4		37001_at				
5		37029_at				
6		37046_at				
7		37189_at				
8		37725_at				
9		38730_at				
10		39425_at				
11		40515_at				
12		41237_at				
13		93822_at				
14		94766_at				
15						
16						

6 22 This is the spreadsheet that lets you enter and edit your data.

Notice that the **Genes** column has been automatically filled with the names of the most significant genes. Information about the T-test is given in the report. These genes will be used later for hierarchical clustering.

T-Test Output



```
T-Test Detail in Probability Level Order
Alternative Hypothesis: Mean of Control - Mean of Treatment <> 0

                                  FDR
                             Adjusted
                             Multiple       Single
Gene         Subset             Tests         Test                Counts        Mean   Standard
Name         Name          Prob Level   Prob Level   T Value     (N1/N2)  Difference      Error
41237_at     Other          0.0001855    0.0000005    57.770         3/3      4.4944     0.0778
101482_at    Other          0.0001912    0.0000011   -48.203         3/3     -3.8749     0.0804
94766_at     Other          0.0005509    0.0000048   -33.403         3/3     -2.8402     0.0850
37001_at     Other          0.0011430    0.0000133   -25.876         3/3     -3.5477     0.1371
37046_at     Other          0.0013776    0.0000200   -23.342         3/3     -3.1501     0.1350
31962_at     Other          0.0016171    0.0000281   -21.414         3/3     -3.8857     0.1815
37029_at     Other          0.0023416    0.0000475   -18.763         3/3     -4.4225     0.2357
38730_at     Other          0.0023667    0.0000549   -18.092         3/3     -3.6455     0.2015
100084_at    Other          0.0070727    0.0001845    13.304         3/3      4.8340     0.3633
93822_at     Other          0.0072033    0.0002088    12.892         3/3      3.8257     0.2968
40515_at     Other          0.0093642    0.0002986    11.766         3/3      4.6568     0.3958
37725_at     Other          0.0138263    0.0004809    10.410         3/3      4.2110     0.4045
39425_at     Other          0.0211671    0.0007976    -9.133         3/3     -3.3847     0.3706
37189_at     Other          0.0445618    0.0018083     7.368         3/3      3.7324     0.5066

Total number of hypothesis tests conducted = 345

Histograms and Plots Section
```

The basic report gives you detailed list of the most significant genes, along with histograms and a volcano plot showing the selected probability level cutoff of 0.05. Notice that the list of genes listed in the **Genes** column on the spreadsheet is the same as the list in this report. There are 14 genes that are differentially expressed, based on this analysis.

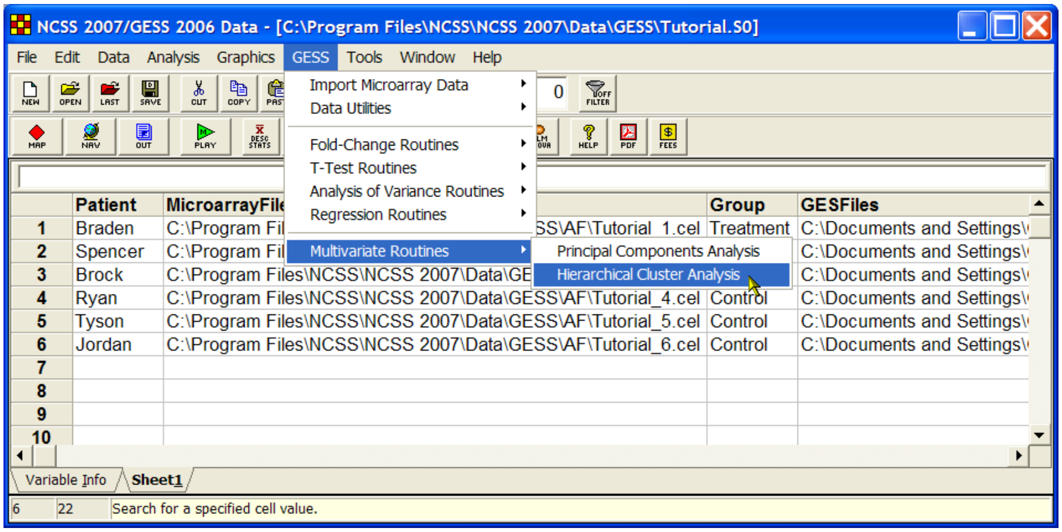
These reports and plots can be copied and pasted directly into a report or presentation. Double-click on any plot to enlarge it.

Hierarchical Cluster Analysis Steps

To run this example, take the following steps or load the **GESS Tutorial - Step 3** template on the Hierarchical Cluster Analysis Template tab.

1 Launch the Hierarchical Cluster Analysis Procedure.

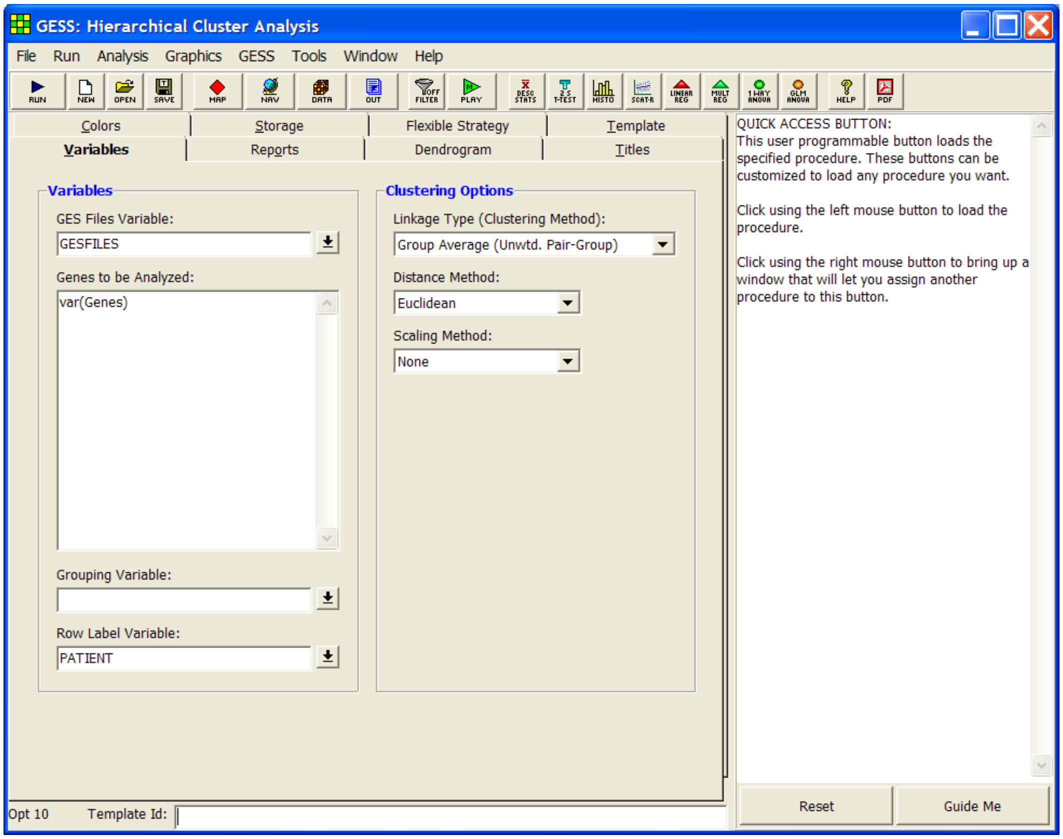
- On the menus, select **GESS**, then **Multivariate Routines**, then **Hierarchical Cluster Analysis**. The Hierarchical Cluster Analysis procedure will be displayed. Hint: Right click on one of the eight quick-launch icons on the spreadsheet toolbar to select icons that you use most frequently. The new icons will be placed on the toolbar.
- On the Hierarchical Cluster Analysis window menus, select **File**, then **New Template**. This will fill the procedure with the default template.



	Patient	MicroarrayFile	Group	GESFiles
1	Braden	C:\Program Fil ... SS\AF\Tutorial_1.cel	Treatment	C:\Documents and Settings\
2	Spencer	C:\Program Fil		C:\Documents and Settings\
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GE		C:\Documents and Settings\
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control	C:\Documents and Settings\
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control	C:\Documents and Settings\
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control	C:\Documents and Settings\
7				
8				
9				
10				

2 Specify the Variables.

- On the Hierarchical Cluster Analysis window, select the **Variables tab**.
- Set the **GES Files Variable** to **GESFiles**.
- Under **Genes to be Analyzed** enter **var(Genes)**.
- Set the **Row Label Variable** to **Patient**.
- Leave all other options under the Variables tab and other tabs at their default settings.


GESS: Hierarchical Cluster Analysis
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
Colors Storage Flexible Strategy Template
Variables Reports Dendrogram Titles
Reset
Guide Me

Variables

GES Files Variable: GESFILES

Genes to be Analyzed: var(Genes)

Grouping Variable:

Row Label Variable: PATIENT

Clustering Options

Linkage Type (Clustering Method): Group Average (Unwtd. Pair-Group)

Distance Method: Euclidean

Scaling Method: None

QUICK ACCESS BUTTON:
This user programmable button loads the specified procedure. These buttons can be customized to load any procedure you want.

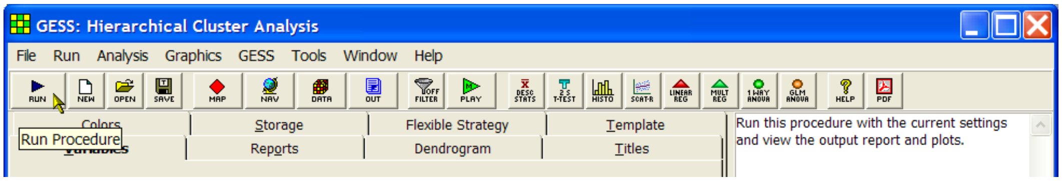
Click using the left mouse button to load the procedure.

Click using the right mouse button to bring up a window that will let you assign another procedure to this button.

Opt 10 Template Id:

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the toolbar at the top).

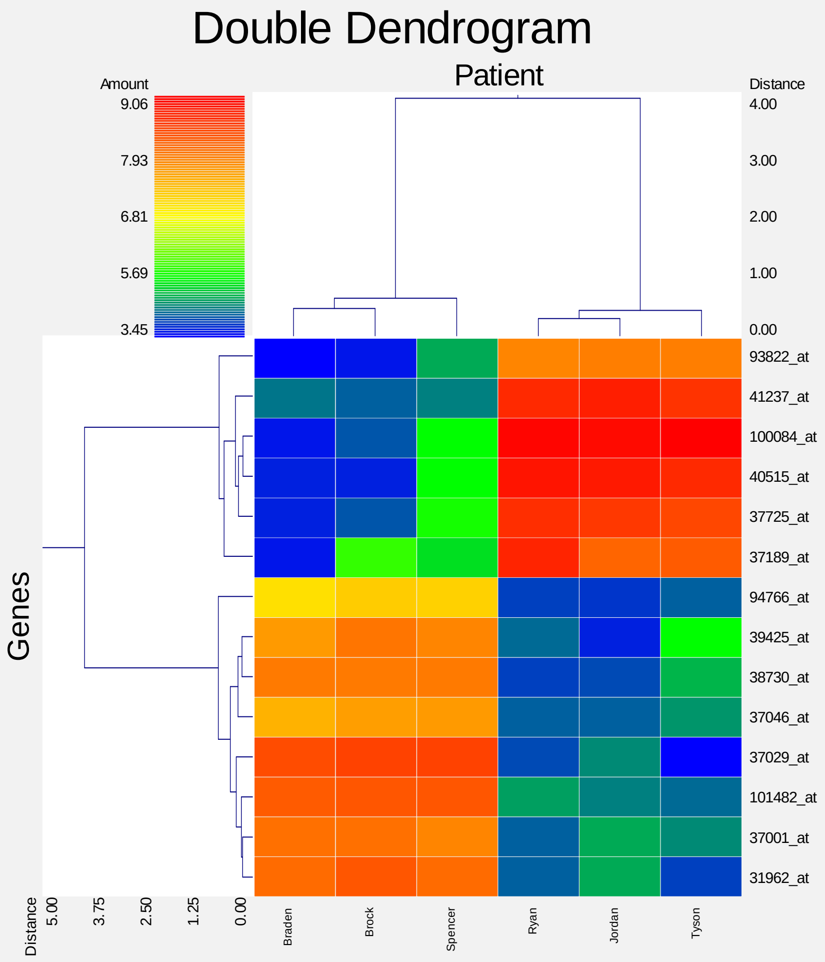

GESS: Hierarchical Cluster Analysis
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
Colors Storage Flexible Strategy Template
Variables Reports Dendrogram Titles
Run Procedure

Run this procedure with the current settings and view the output report and plots.

Hierarchical Clustering Output

```
Clustering Method   Group Average (Unweighted Pair-Group)
Distance Type       Euclidean
Scale Type          None

Cluster Detail Section when Clustering Rows

Cluster      Rows in this Cluster
1            Braden, Brock
2            Ryan, Tyson, Jordan
None         Spencer

Cluster Detail Section when Clustering Genes

Cluster      Genes in Cluster
1            31962_at, 37001_at, 37029_at, 37046_at, 38730_at, 39425_at, 101482_at
2            37189_at, 37725_at, 40515_at, 41237_at, 100084_at, 93822_at
None         94766_at
```

Dendrogram Section



The basic report gives you a list of the clusters and a double dendrogram. The apparent separation in patient clusters corresponds exactly with the two treatment groups.

These reports and plots can be copied and pasted directly into a report or presentation. Double-click on the double dendrogram to enlarge it.

Chapter 100

Installation and Basics

Before You Install

1. Check System Requirements

GESS runs on 32-bit and 64-bit Windows systems. This includes Windows ME, Windows NT 4.0, Windows 2000, Windows XP, and Windows Vista. The recommended minimum system is a Windows XP or Vista-compatible PC.

GESS takes up about 53 MB of disk space. Once installed, ***GESS*** also requires about 20 MB of temporary disk space while it is running. ***NCSS*** must be installed to run ***GESS***, but a license for ***NCSS*** is not required. ***NCSS*** requires about 120 MB of disk space.

2. Find a Home for GESS

Before you start installing, decide on a directory where you want to install ***NCSS*** and ***GESS***. By default, the setup program will install ***NCSS*** and ***GESS*** in the *C:\Program Files\NCSS\NCSS 2007* directory. You may change this during the installation, but not after. The example data, template, and macro files will be placed in your personal documents folder (usually *C:\…\[My] Documents\NCSS\NCSS 2007*) in appropriate subdirectories. The program will save all procedure templates and macros to these folders while the program is running.

3. If You Already Own GESS, NCSS, or PASS

GESS and ***NCSS*** have been combined into one physical program. Access to each program is controlled by separate serial numbers. If you have a serial number for ***GESS***, but not for ***NCSS***, ***NCSS*** will work as trial versions for 7 days from the time the first procedure is accessed. Some of the ***NCSS*** routines (Scatter Plots, Box Plots, Histograms, and Data Report) are included with your purchase of ***GESS***.

If ***NCSS*** or ***GESS*** is already installed on your system, instruct the installation program to place this new version in a new folder (e.g. *C:\Program Files\NCSS\NCSS 2007*). All appropriate files will be copied from your old ***NCSS*** directory or replaced by updated files.

What Install Does

The installation procedure creates the necessary folders and copies the ***NCSS/GESS*** program from the installation file, called *NCSS2007SETUP.EXE*, to those folders. The files in

NCSS2007SETUP.EXE are compressed, so the installation program decompresses these files as it copies them to your hard disk. The following folders are created during installation:

C:\Program Files\NCSS\NCSS 2007 (or your substitute folder) contains most of the program files.

C:\Program Files\NCSS\NCSS 2007\Data\GESS contains data files required for ***GESS*** tutorials.

C:\Program Files\NCSS\NCSS 2007\Pdf\GESS contains printable copies of the documentation in PDF format.

C:\Program Files\NCSS\NCSS 2007\Sts contains all labels, text, and online messages.

C:\…\[My] Documents\NCSS\NCSS 2007\Data\GESS contains the database files used by the tutorials. We recommend creating a sub-folder of this folder to contain the data for each project you work on.

C:\…\[My] Documents\NCSS\NCSS 2007\Junk contains temporary files used by the program while it is running. Under normal operation, ***GESS*** will automatically delete temporary files. After finishing ***GESS***, you can delete any files left in this folder (but not the folder itself).

C:\…\[My] Documents\NCSS\NCSS 2007\Macros contains saved macros.

C:\…\[My] Documents\NCSS\NCSS 2007\Report is the default folder in which to save your output. You can save the reports to any folder you wish.

C:\…\[My] Documents\NCSS\NCSS 2007\Settings contains the files used to store your template files. These files are used by the ***GESS*** template system, which is described in a later chapter.

Installing NCSS and GESS

This section gives instructions for installing ***NCSS*** and ***GESS*** on your computer system. You must use the ***NCSS/GESS*** setup program to install ***GESS***. The files are compressed, so you cannot simply copy the files to your hard drive.

Follow these basic steps to install ***NCSS*** and ***GESS*** on your computer system.

Step	Notes
1.	Make sure that you are using a 32- or 64-bit version of windows such as Windows Me, Windows NT 4.0, Windows 2000, Windows XP, or Windows Vista.
2.	If you are installing from a CD, insert the CD in the CD drive. The installation program should start automatically. If it does not, on the Start menu, select the Run command. Enter *D:\NCSS\NCSS2007Setup*. You may have to substitute the appropriate letter for your CD drive if it is not *D*. If you are installing from a download, simply run the downloaded file (*NCSS2007Setup.exe*).
3.	Once the setup starts, follow the instructions on the screen. ***NCSS*** and ***GESS*** will be installed in the drive and folder you designate.

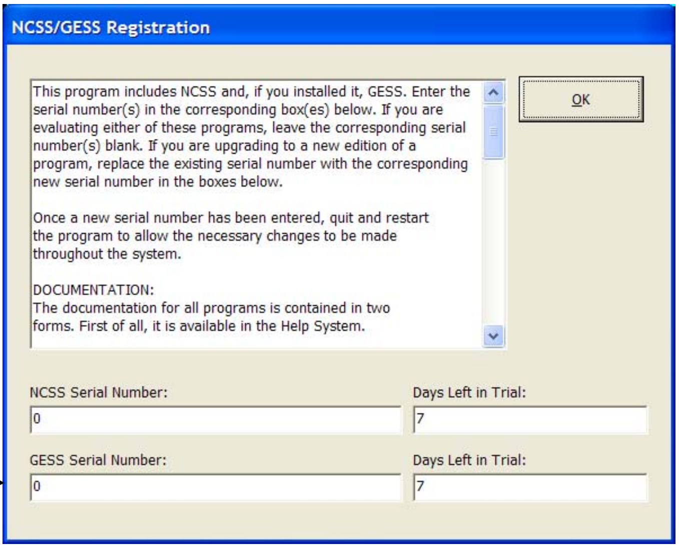
If Something Goes Wrong During Installation

The installation procedure is automatic. If something goes wrong during installation, delete the \NCSS2007 directory and start the installation process from the beginning. If trouble persists, contact our technical support staff as indicated below.

Starting GESS

GESS may be started using your keyboard or your mouse using the same techniques that you use to start any other Windows application. You can start ***GESS*** by selecting **NCSS 2007** from your Start menu using standard mouse or keyboard operations.

The first time you run ***GESS***, enter your serial number in the pop-up window that appears when the program begins. If you also have a serial number for ***NCSS***, enter it in the corresponding window. If you do not enter a serial number, ***NCSS*** and/or ***GESS*** will enter trial mode and you will have 7 days to evaluate ***NCSS*** and ***GESS***. When in trial mode, ***NCSS*** and ***GESS*** are fully-functional but limited to 100 rows of data.



NCSS/GESS Registration

This program includes NCSS and, if you installed it, GESS. Enter the serial number(s) in the corresponding box(es) below. If you are evaluating either of these programs, leave the corresponding serial number(s) blank. If you are upgrading to a new edition of a program, replace the existing serial number with the corresponding new serial number in the boxes below.

Once a new serial number has been entered, quit and restart the program to allow the necessary changes to be made throughout the system.

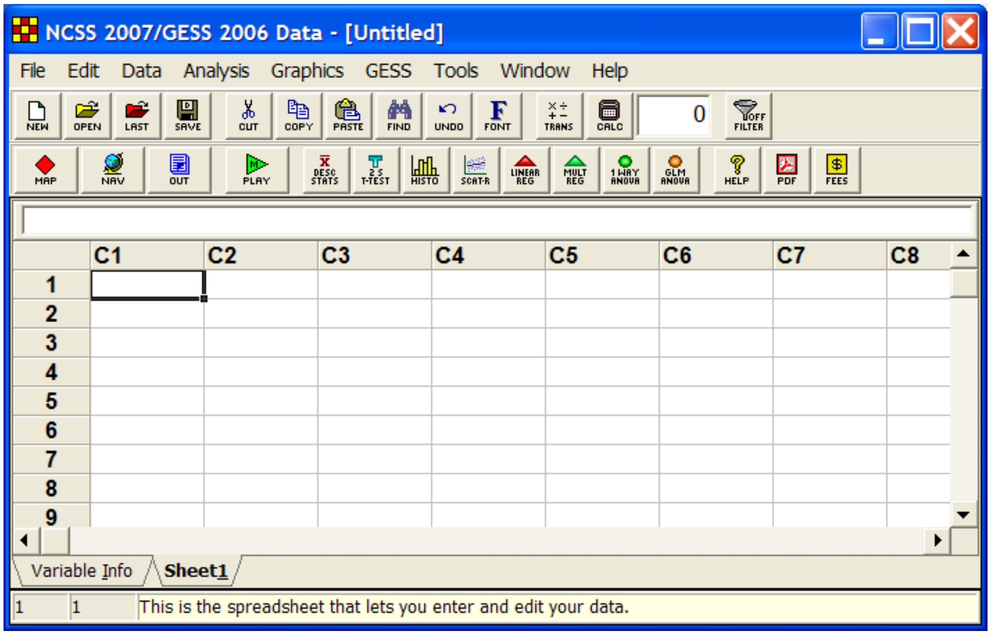
DOCUMENTATION:
The documentation for all programs is contained in two forms. First of all, it is available in the Help System.

NCSS Serial Number: 0
Days Left in Trial: 7

GESS Serial Number: 0
Days Left in Trial: 7

Enter your ***GESS*** serial number here.

The ***NCSS/GESS Data*** window will appear.

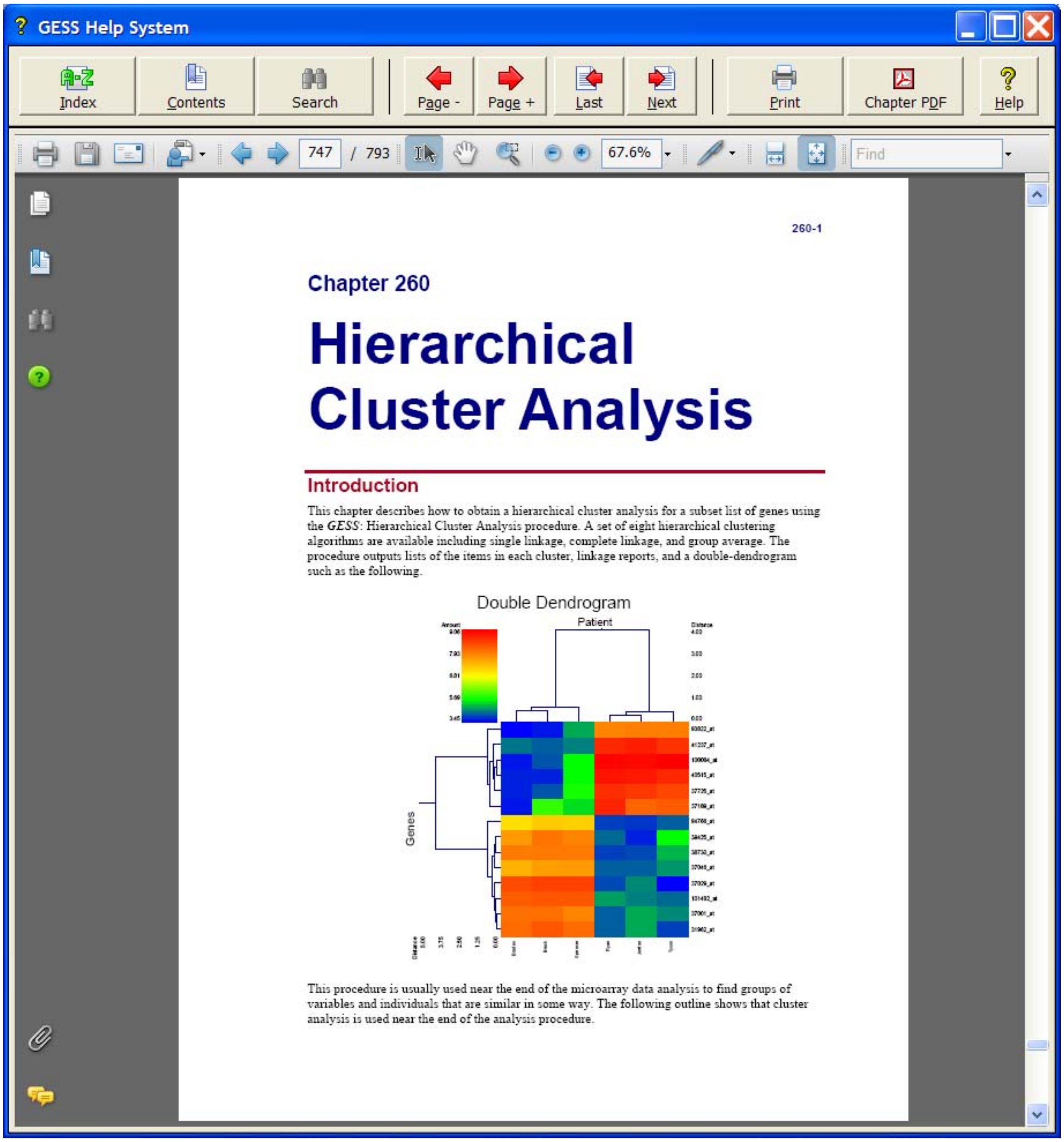


This is the spreadsheet that lets you enter and edit your data.

Obtaining Help

Help System

To help you learn and use ***GESS*** efficiently, the material in this manual is included in the ***GESS*** Help System. The Help System is started from the Help menu or by clicking on the yellow '?' icon on the right side of the toolbar. ***GESS*** updates, available for download at www.ncss.com, may contain adjustments or improvements of the ***GESS*** Help System. Adobe Acrobat or Adobe Reader version 7 or 8 is required to view the help system. You can download Adobe Reader 8 for free by going to www.adobe.com. Adobe Reader 8 can also be installed from the *Utilities* folder on your ***GESS*** installation CD.



260-1

Chapter 260

Hierarchical Cluster Analysis

Introduction

This chapter describes how to obtain a hierarchical cluster analysis for a subset list of genes using the *GESS*: Hierarchical Cluster Analysis procedure. A set of eight hierarchical clustering algorithms are available including single linkage, complete linkage, and group average. The procedure outputs lists of the items in each cluster, linkage reports, and a double-dendrogram such as the following.

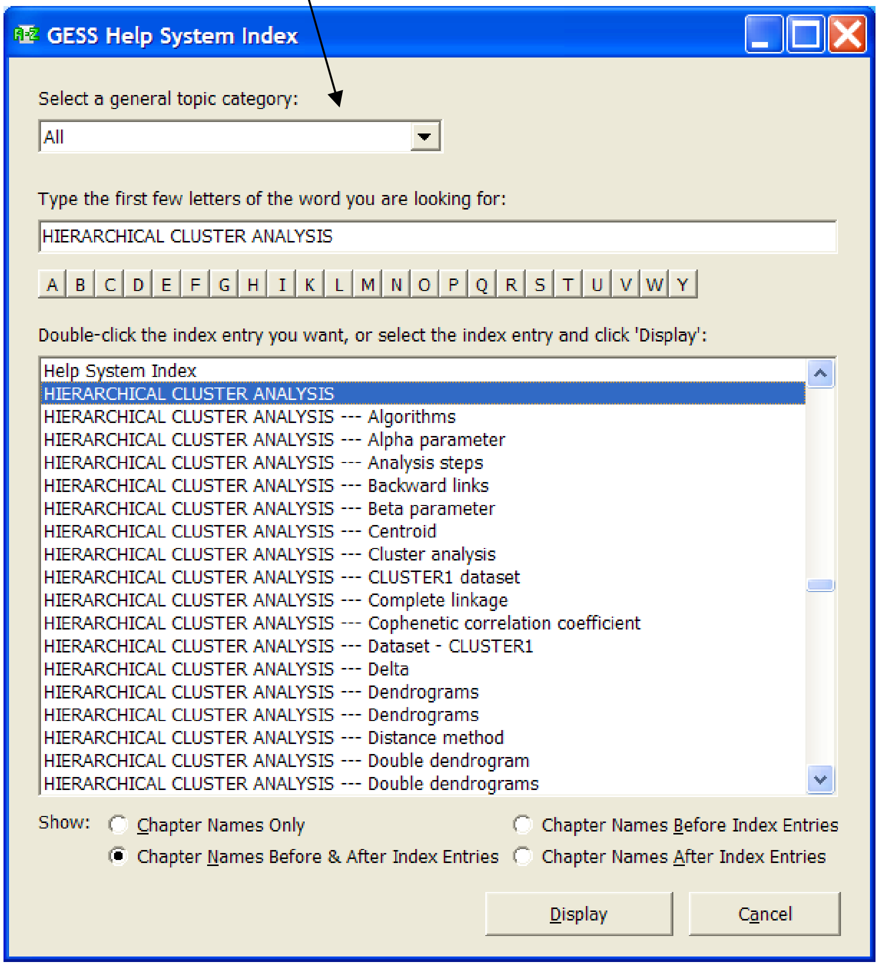
This procedure is usually used near the end of the microarray data analysis to find groups of variables and individuals that are similar in some way. The following outline shows that cluster analysis is used near the end of the analysis procedure.

Navigating the Help System

There are a few key features of our help system that will let you use the help system more efficiently. We will now explain each of these features.

Index Window

The Index Window can be launched at any time by clicking on the Index button on the ***GESS*** Help System display window. The index allows you to quickly locate keywords and/or statistical topics. You can narrow the list of index entries displayed by selecting a specific topic category in the uppermost dropdown box.



GESS Help System Index

Select a general topic category: All

Type the first few letters of the word you are looking for: HIERARCHICAL CLUSTER ANALYSIS

Double-click the index entry you want, or select the index entry and click 'Display':

Help System Index
HIERARCHICAL CLUSTER ANALYSIS
HIERARCHICAL CLUSTER ANALYSIS --- Algorithms
HIERARCHICAL CLUSTER ANALYSIS --- Alpha parameter
HIERARCHICAL CLUSTER ANALYSIS --- Analysis steps
HIERARCHICAL CLUSTER ANALYSIS --- Backward links
HIERARCHICAL CLUSTER ANALYSIS --- Beta parameter
HIERARCHICAL CLUSTER ANALYSIS --- Centroid
HIERARCHICAL CLUSTER ANALYSIS --- Cluster analysis
HIERARCHICAL CLUSTER ANALYSIS --- CLUSTER1 dataset
HIERARCHICAL CLUSTER ANALYSIS --- Complete linkage
HIERARCHICAL CLUSTER ANALYSIS --- Cophenetic correlation coefficient
HIERARCHICAL CLUSTER ANALYSIS --- Dataset - CLUSTER1
HIERARCHICAL CLUSTER ANALYSIS --- Delta
HIERARCHICAL CLUSTER ANALYSIS --- Dendrograms
HIERARCHICAL CLUSTER ANALYSIS --- Dendrograms
HIERARCHICAL CLUSTER ANALYSIS --- Distance method
HIERARCHICAL CLUSTER ANALYSIS --- Double dendrogram
HIERARCHICAL CLUSTER ANALYSIS --- Double dendrograms

Show:
○ Chapter Names Only ○ Chapter Names Before Index Entries
● Chapter Names Before & After Index Entries ○ Chapter Names After Index Entries

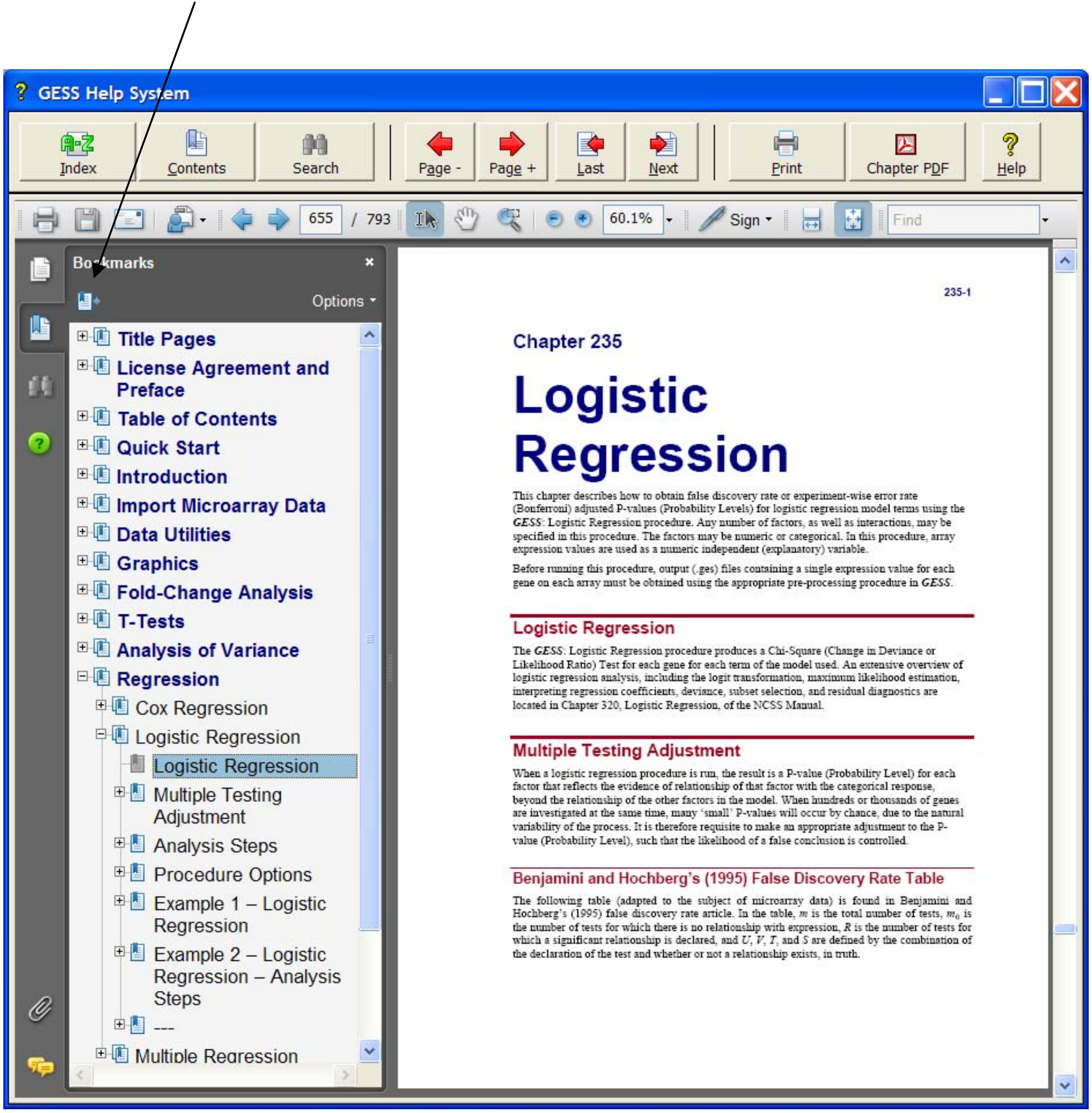
Index entries are displayed in the format

Index Entry --- CHAPTER NAME or CHAPTER NAME --- Index Entry.

You can control which entries are displayed by clicking on the radio buttons at the bottom of the window.

Contents Window

Clicking on the Contents button opens the Contents (Bookmarks) Window of the viewer. From this window you can expand the table of contents to view nested headings. You can click on the “Expand Current Bookmark” icon to instantly find the bookmark location for the currently-displayed page in the help document.



Chapter 235

Logistic Regression

This chapter describes how to obtain false discovery rate or experiment-wise error rate (Bonferroni) adjusted P-values (Probability Levels) for logistic regression model terms using the *GESS*: Logistic Regression procedure. Any number of factors, as well as interactions, may be specified in this procedure. The factors may be numeric or categorical. In this procedure, array expression values are used as a numeric independent (explanatory) variable.

Before running this procedure, output (.ges) files containing a single expression value for each gene on each array must be obtained using the appropriate pre-processing procedure in *GESS*.

Logistic Regression

The *GESS*: Logistic Regression procedure produces a Chi-Square (Change in Deviance or Likelihood Ratio) Test for each gene for each term of the model used. An extensive overview of logistic regression analysis, including the logit transformation, maximum likelihood estimation, interpreting regression coefficients, deviance, subset selection, and residual diagnostics are located in Chapter 320, Logistic Regression, of the NCSS Manual.

Multiple Testing Adjustment

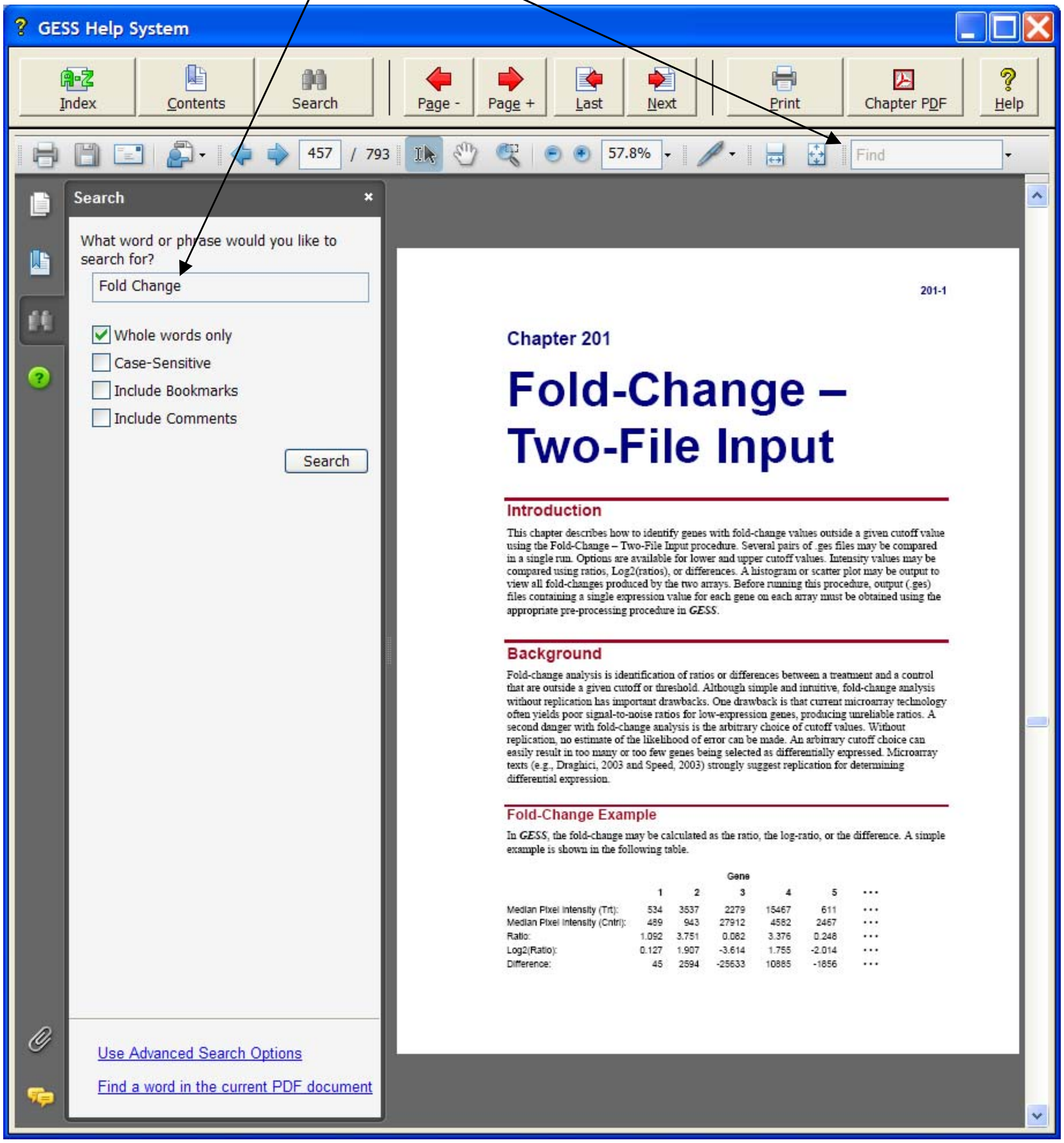
When a logistic regression procedure is run, the result is a P-value (Probability Level) for each factor that reflects the evidence of relationship of that factor with the categorical response, beyond the relationship of the other factors in the model. When hundreds or thousands of genes are investigated at the same time, many ‘small’ P-values will occur by chance, due to the natural variability of the process. It is therefore requisite to make an appropriate adjustment to the P-value (Probability Level), such that the likelihood of a false conclusion is controlled.

Benjamini and Hochberg’s (1995) False Discovery Rate Table

The following table (adapted to the subject of microarray data) is found in Benjamini and Hochberg’s (1995) false discovery rate article. In the table, m is the total number of tests, m_0 is the number of tests for which there is no relationship with expression, R is the number of tests for which a significant relationship is declared, and U, V, T, and S are defined by the combination of the declaration of the test and whether or not a relationship exists, in truth.

Search Window

Clicking on the Search button opens the Search Window of the viewer. From this window you can search the entire help system for any word or phrase. A search can also be initiated from the Find box in the viewer toolbar.



What word or phrase would you like to search for?

Fold Change

☑ Whole words only

□ Case-Sensitive

□ Include Bookmarks

□ Include Comments

201-1

Chapter 201

Fold-Change – Two-File Input

Introduction

This chapter describes how to identify genes with fold-change values outside a given cutoff value using the Fold-Change – Two-File Input procedure. Several pairs of .ges files may be compared in a single run. Options are available for lower and upper cutoff values. Intensity values may be compared using ratios, Log2(ratios), or differences. A histogram or scatter plot may be output to view all fold-changes produced by the two arrays. Before running this procedure, output (.ges) files containing a single expression value for each gene on each array must be obtained using the appropriate pre-processing procedure in *GESS*.

Background

Fold-change analysis is identification of ratios or differences between a treatment and a control that are outside a given cutoff or threshold. Although simple and intuitive, fold-change analysis without replication has important drawbacks. One drawback is that current microarray technology often yields poor signal-to-noise ratios for low-expression genes, producing unreliable ratios. A second danger with fold-change analysis is the arbitrary choice of cutoff values. Without replication, no estimate of the likelihood of error can be made. An arbitrary cutoff choice can easily result in too many or too few genes being selected as differentially expressed. Microarray texts (e.g., Draghici, 2003 and Speed, 2003) strongly suggest replication for determining differential expression.

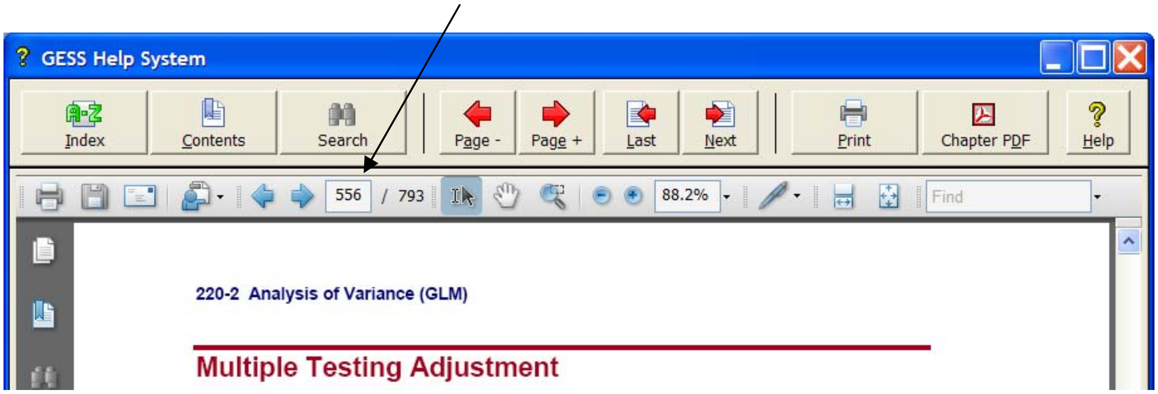
Fold-Change Example

In *GESS*, the fold-change may be calculated as the ratio, the log-ratio, or the difference. A simple example is shown in the following table.

	Gene					
	1	2	3	4	5	• • •
Median Pixel Intensity (Trt):	534	3537	2279	15467	611	• • •
Median Pixel Intensity (Cntrl):	489	943	27912	4582	2467	• • •
Ratio:	1.092	3.751	0.082	3.376	0.248	• • •
Log2(Ratio):	0.127	1.907	-3.614	1.755	-2.014	• • •
Difference:	45	2594	-25633	10885	-1856	• • •

Printing the Documentation

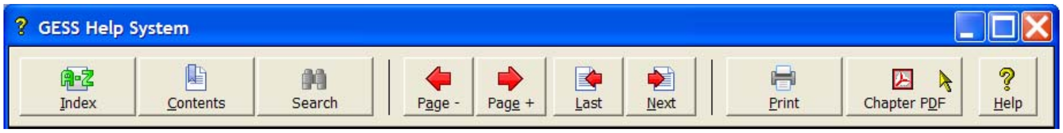
To print pages from the documentation, click on the **Print** button on the ***GESS*** Help System toolbar. This will launch the Adobe Reader print dialogue screen. You can choose to print a single page or a range of pages from the help file. When entering page numbers, remember to use the PDF file page numbers (e.g., 556-558) and not the page numbers found in the document pages (e.g., 220-2 to 220-4 is not a valid page range). The Adobe Reader page numbers can be seen in the viewer window.



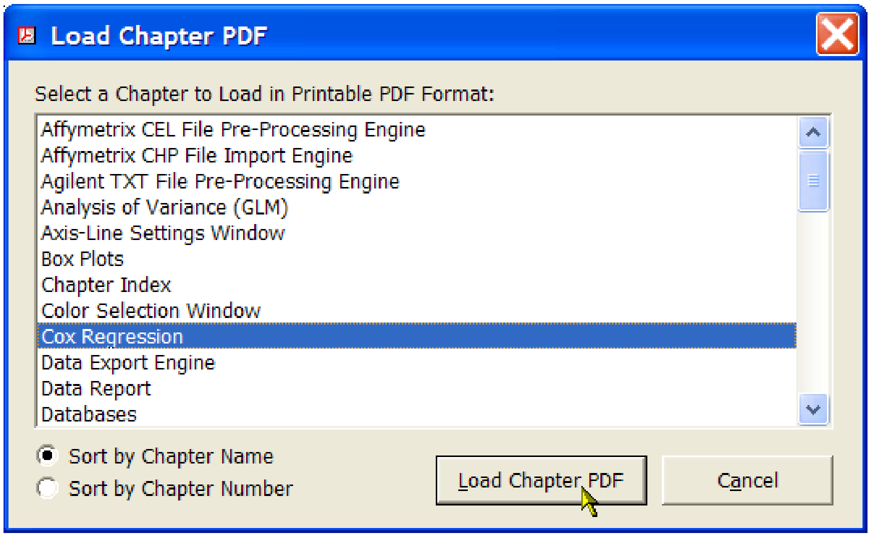
If you are using Adobe Reader 7, then the page numbers are found at the bottom of the viewer window.

To print a single chapter or topic using your default PDF viewer, take the following steps:

1. Click on the **Chapter PDF** icon in the ***NCSS*** Help System toolbar.



2. Choose the chapter you would like to print from the list and click **Load Chapter PDF**. This will launch the individual chapter PDF in a separate window using your default PDF viewer (e.g., Adobe Reader).



Load Chapter PDF

Select a Chapter to Load in Printable PDF Format:

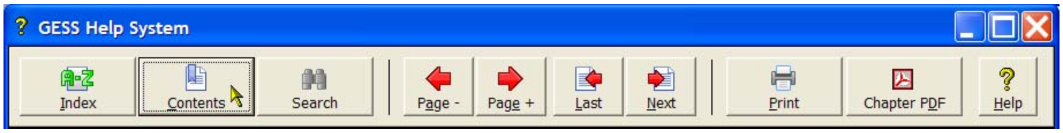
Affymetrix CEL File Pre-Processing Engine
Affymetrix CHP File Import Engine
Agilent TXT File Pre-Processing Engine
Analysis of Variance (GLM)
Axis-Line Settings Window
Box Plots
Chapter Index
Color Selection Window
Cox Regression
Data Export Engine
Data Report
Databases

● Sort by Chapter Name
○ Sort by Chapter Number

3. Use the **Print** function of your PDF viewer to print the entire chapter or individual pages from the chapter.

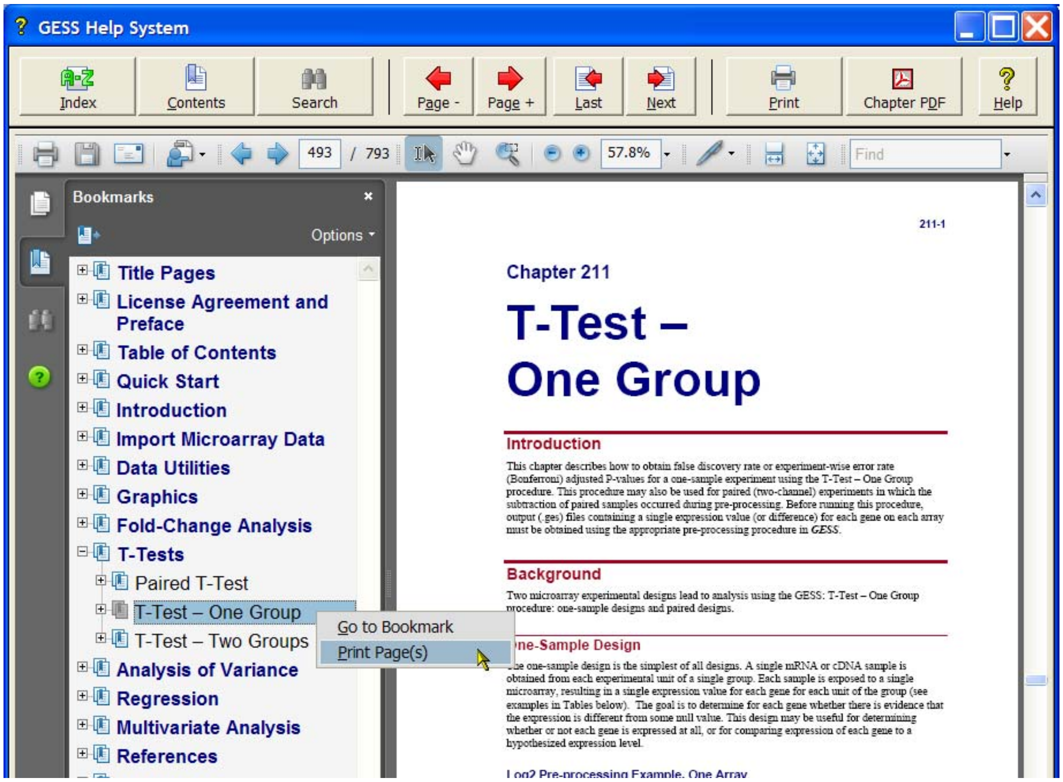
If you have Adobe Reader 8 or later, you can print entire chapters using an alternative method as follows (**This will not work with Adobe Reader 7**):

1. Open the Contents (Bookmarks) Window by clicking on the **Contents button** at the top of the ***GESS*** Help System display window.



2. Expand the bookmarks to display the chapter or topic name you wish to print (e.g., the T-Test – One Group chapter). Then, **highlight** the chapter name, **right-click** on the highlighted selection (or select Options in the panel above), and select **Print Page(s)**. This will automatically print only the pages from the selected chapter.

   CAUTION: When you click Print Page(s), the command is sent to the printer automatically without any intermediate Print Setup window being displayed. Make sure that you have selected only the topic you want before clicking Print Page(s).



211-1

Chapter 211

T-Test – One Group

Introduction

This chapter describes how to obtain false discovery rate or experiment-wise error rate (Bonferroni) adjusted P-values for a one-sample experiment using the T-Test – One Group procedure. This procedure may also be used for paired (two-channel) experiments in which the subtraction of paired samples occurred during pre-processing. Before running this procedure, output (.ges) files containing a single expression value (or difference) for each gene on each array must be obtained using the appropriate pre-processing procedure in *GESS*.

Background

Two microarray experimental designs lead to analysis using the GESS: T-Test – One Group procedure: one-sample designs and paired designs.

ne-Sample Design

e one-sample design is the simplest of all designs. A single mRNA or cDNA sample is obtained from each experimental unit of a single group. Each sample is exposed to a single microarray, resulting in a single expression value for each gene for each unit of the group (see examples in Tables below). The goal is to determine for each gene whether there is evidence that the expression is different from some null value. This design may be useful for determining whether or not each gene is expressed at all, or for comparing expression of each gene to a hypothesized expression level.

Log2 Pre-processing Example, One Array

If you do not want to print the entire chapter, continue to expand the bookmark tree to the topic you wish to print before completing step 2. The Print Page(s) command prints all pages containing bookmarks that are nested within the highlighted bookmark.

Technical Support

If you have a question about ***GESS***, you should first look to the printed documentation and the included Help system. If you cannot find the answer there, look for help on the web at www.ncss.com/support.html. If you are unable to find the answer to your question by these means, contact ***NCSS*** technical support for assistance by calling (801) 546-0445 between 8 a.m. and 5 p.m. (MST). You can contact us by email at support@ncss.com or by fax at (801) 546-3907. Our technical support staff will help you with your question.

If you encounter problems or errors while using ***GESS***, please view our list of recent corrections before calling by going to www.ncss.com/release_notes.html to find out if you problem or error has been corrected by an update. You can download updates anytime by going to www.ncss.com/download.html. If updating your software does not correct the problem, contact us by phone or email.

To help us answer your questions more accurately, we may need to know about your computer system. Please have pertinent information about your computer and operating system available.

Chapter 101

Databases

Introduction

GESS analyzes data contained in a database. There are two types of ***GESS*** databases: S0 and S0Z. We will now explain these two types of databases to you.

S0 (Spreadsheet) Database

An S0 database (sometimes referred to as a worksheet) is made up of a Variable-Info sheet and one or more datasheets. Variable names, labels, formats, and transformations are contained on the Variable-Info sheet. Each datasheet contains 256 variables (columns) and room for up to 16,384 observations, although we recommend that you use this type for databases with less than 1000 observations. You can add as many datasheets to a database as you like (thereby increasing the number of variables), but you cannot increase the number of observations.

When we refer to a variable on a ***GESS*** database, we are actually referring to a specific column of a datasheet. All procedures analyze one or more variables from these datasheets.

GESS accepts both numeric and text data. Numeric data may contain up to 16 digits (double-precision). Text data may contain up to 1000 characters per cell.

Physically, an S0-type database is made up of two or more files with appropriate extensions. The Variable-Info file has the extension S0. The datasheets have the extensions S1, S2, etc. Hence, a database called “ABC” with 512 variables (two datasheets) would appear on your hard drive as three files: ABC.S0, ABC.S1, and ABC.S2. This is important to remember when backing up or copying an S0-type database.

Each of these files is actually a *Microsoft Excel 4.0* compatible spreadsheet file. This is where the row and column limits come from since an Excel 4.0 spreadsheet can contain up to 256 columns and 16,384 rows. We have used this format because it is popular, transportable, and because it allows us to provide a familiar, spreadsheet-style interface complete with formatting and formulas.

S0Z (Zipped) Database

The S0Z-type database is disk-based: only small portions of the database reside in your computer’s memory at any one time. Because this type of database is not memory resident, it can be much larger: up to about 5,000 variables and over 100,000 rows. Because the complete database is not stored in your computer’s RAM memory, it is slower to access and the “Undo” feature is not available. On the positive side, because all data is not loaded into memory at one time, it does not require as much RAM memory to run medium (100 - 500 rows) to large (10000+ rows) databases.

Unlike most commercial databases, variables (fields) in S0Z-type databases can contain a mixture of numeric and text values. You do not have to specify the field type in advance. Numeric data may contain up to 16 digits (double-precision). Text data may contain as many characters as you have made space for on a data record. You define the record size when the database is initialized.

Technical Details of S0Z Databases

An understanding of how this database is constructed will help you make intelligent choices when you create one. The S0Z database may be viewed as one long row of data which is made up of chunks called “records.” Each record contains a fixed number of characters (bytes), which cannot be changed (without copying the existing data to a new database with a larger record length).

Record 1

The first record contains database-size information such as the number of variables, number of rows, and record length. Its length is 512 bytes.

Records 2 through M+1

The next M records (where M represents the number of variables) contain information about the variables. Each record contains information about the corresponding variable, such as its name, label, transformation, format, and data type.

Records M+2 to M+N+1

The next N records (where N represents the number of observations or rows) contain the data values. Each record holds one row of data. The record is made up of M blocks of k bytes each followed by T blocks of m bytes each.

The first M blocks hold the regular data values, one per block. Numeric values (including date) use the first 8 bytes of a block. Text values use the last k-4 bytes of a block. Hence, the default, 10-byte, block can hold a text value of up to 6 bytes.

The last T blocks provide additional storage for extra long text values. When a text value will not fit in the regular data block, it is stored in the first available text block at the end of the record. If all of these overflow blocks are full, the text is truncated to k-4 bytes.

When you create a database, you must pay particular attention to how much space you want to provide for text data. You have two alternatives. First, you can increase the size of k so that most of your text values will be stored in the first M blocks. This method increases the overall size of the database, but provides the fastest access. Second, you can increase T so that there are more overflow blocks. This method provides slightly slower access to your data, but makes for a smaller database.

Note that if you are planning to use value labels, you need to include enough room to hold them.

Calculating Record Size

The record size is calculated as

$$L = Mk + Tm$$

where

L	Record size in bytes
M	Number of variables
k	Length of a variable field (default is 10 bytes)
T	Number of extra text fields
m	Length of an extra text field (default is 25 bytes)

For example, suppose you want to create a database that will have space for 50 numeric variables and 10 text variables. Suppose the largest text field is 30 characters long. The record size would be

$$50(10)+10(30) = 800 \text{ bytes.}$$

As explained next, you should be liberal in your assignment of the maximum number of variables. Make sure you include enough space for new and transformed variables. Also, be sure you include enough text variables to hold all value labels that you might use.

Zipped Database

To provide flexibility in what may be stored, we suggest that you always add a little extra to the number of bytes you think will be required. Because this will obviously “waste” some disk space, we have added one final operation: your database is permanently stored in a compressed (zipped) format. Because of this, you do not have to be so careful about how much extra space you add to each record. When the file is compressed, it is often reduced to one-tenth the size. For example, the size of our SAMPLE database (50 columns by 75 rows) is about 100,000 bytes. When compressed, its size is only 12,000 bytes!

When you load a S0Z database, the actual file is not “loaded.” Instead, an uncompressed copy of the database (with the extension “S0N”) is created in your temporary NCSS directory. As you read and write to and from the database, you are actually using the temporary database, not the original.

When you save the database, the program compresses the temporary database and replaces the existing S0Z file with the new compressed version. This means that you can always “undo” operations that have resulted in your data being unintentionally modified. You simply re-open your database without saving the modifications. When the program asks if you want to save changes, just say no! The modified temporary file will be replaced by another expanded copy of your compressed database.

Comparison of S0 and S0Z Databases

The following table presents a brief comparison of these two types of databases. It will help you determine which type to use in a particular situation.

Criterion	S0 Database	S0Z Database
Rows	Recommend < 1000	Unlimited
Columns	Recommend < 50	Up to 16,000
Transformations	Yes	Yes
Cut/Copy/Paste	Yes	Yes
Insert/Delete	Yes	Yes
Undo	Yes	No
Speed	Fast	Medium
File Size	Regular	Small (Compressed)
Text Data	Yes	Yes
Numeric Data	Yes	Yes
Flexibility	Yes	Must stay within limits set when created.

Accuracy

GESS maintains double-precision (sixteen digit) accuracy in the data values. When you apply a special format to a variable, only the display of the data is modified. The actual sixteen-digit number is maintained on the datasheet.

All numbers are stored in the IEEE floating-point format. Such numbers range from 4.94E-324 to 1.798E308 for positive values, 0, and from -1.798E308 to -4.940E-324 for negative values.

Because of the rounding that has to occur to fit into the IEEE format, it is impossible to express all numbers exactly. For example, the common fraction of one-third cannot be expressed exactly, but must be rounded at the last digit. This rounding may cause strange results for certain decimal numbers. For example, you might enter 453.4537 and have it displayed as 453.45369999999999. There is nothing we can do to change this oddity. You can rescale your data by multiplying the variable by 10 or 100. However, this problem occurs rarely in practice and will not change the accuracy of your results.

The “E” notation is used for expressing large and small numbers in scientific notation. The number after the E is the exponent (base 10) that is applied to the base number. Thus, 3.238E-03 means 0.003238. Other examples are:

E-Notation	Decimal Equivalent
-3.238E-06	-.000003238
-3.238E-02	-.03238
3.238E04	32380.0
3.238E06	3238000.0

Missing Values

Missing values are represented by empty cells. In procedures that use only numeric values, such as the average, text values are also treated as missing values. Unless otherwise specified, ***GESS*** uses *rowwise deletion* of observations with missing values. This means that if any of the active variables in a row has a missing value, the whole observation is omitted from the calculations.

Variable Info

Each column of a datasheet is called a *variable*. Each S0 datasheet contains space for 256 variables. Hence, if you have a survey with 700 questions, you will need three datasheets in your database to hold the data, or you could use one S0Z database.

You should note that even though a datasheet has space for 256 variables, only cells that actually contain data are stored when you save the database. You will not be wasting a lot of disk space when you use only 5 or 10 of the 256 variables that are available on the datasheet.

All aspects of a variable (except for the actual data) are modified on the Variable Info sheet. You can view the Variable Info sheet by clicking the Variable Info tab at the bottom of the spreadsheet.

The Variable Info may be printed out using the Print option of the File Menu.

Note that you can modify the information on the Variable Info datasheet without actually viewing it. This may be accomplished by selecting Variable Info from the Edit menu. The Edit menu may be activated using the standard menu at the top of the screen or by clicking the right-mouse button while the mouse pointer is over a specific variable on the datasheet.

We will now explain each column of the Variable Info sheet. Note that each row of this sheet refers to a specific variable on the database.

Number

The first column of the Variable Info sheet gives the variable numbers. A variable’s number is determined by its position on the database. Although you refer to a variable by its name when you write transformations and make variable selections, your selection may be stored in the Procedure Template by number. This is all handled internally, so you will not have to worry about it except when you move variables around within a database. When you do this, you will have to check stored procedure templates to make sure they still refer to the correct variables.

Name

The second column of the Variable Info sheet gives the variable names. Each variable is given a default name when the database is created. Throughout the program, you refer to the variable by its name. The default names are C1, C2, C3, etc.

The variable’s name may be changed at any time by editing the Name column of the Variable Info sheet. The syntax for the naming of variables follows these rules:

1. Variable names must begin with a letter.
2. Variable names can contain only letters, numbers, and the underscore.

3. Variable names may not include spaces.
4. Variable names may not include mathematical symbols.
5. Upper- and lower case letters are the same. Names are case insensitive.

Label

The third column of the Variable Info sheet gives the variable labels. Each variable may have a label associated with it. This label is of arbitrary length. No attempt is made to trim this label if it is too long for the space provided to display it in a particular report. Each Procedure Template has an option in which you designate whether to include these labels on the output.

Transformation

The fourth column of the Variable Info sheet holds a variable's transformation, if one has been assigned. Variable transformations are discussed in detail in another section.

Format

The fifth column of the Variable Info sheet gives the variable format. When this value is left blank, the *General* format is assumed. You can enter a separate format statement for each variable. The format controls both the display and color of the data.

A special Format window may be used to modify a variable's format. This window may be viewed by double clicking in the Format column or by selecting Edit, then Variable Info, then Format from the menus. Note that the Edit menu may also be activated using the right-mouse button.

Several built-in formats are available and it is easy to write your own custom format. Each custom format can have as many as four sections, separated by semicolons:

1. One for positive numbers
2. One for negative numbers
3. One for zeros
4. One for text

Each section is optional, but if you have more than one, you define the format section by the placement of extra semicolons. If you only use one format section, it defines the format for all numbers, both positive and negative. Following is an example of a typical, four-section format:

```
0.000;(0.000);0; "Numeric Only"
```

The following table lists the format symbols that can be used in a custom format.

Format Symbol	Description
General	Display a number in general format.
0	This is a digit placeholder. If the number contains fewer digits than the format contains placeholders, the number is padded with 0’s. If there are more digits to the right of the decimal than there are placeholders, the decimal portion is rounded to the number of places specified by the placeholders. If there are more digits to the left of the decimal than there are placeholders, the extra digits are retained.
#	Digit placeholder. This placeholder functions the same as the 0 placeholder except that the number is not padded with 0’s if the number contains fewer digits than the format contains placeholders.
?	Digit placeholder. This placeholder functions the same as the 0 placeholder except that spaces are used to pad the digits.
. (period)	Decimal point. Determines how many digits (0’s or #’s) are displayed on either side of the decimal point. If the format contains only #’s left of the decimal point, numbers less than 1 begin with a decimal point. If the format contains 0’s left of the decimal point, numbers less than 1 begin with a 0 left of the decimal point.
%	Display the number as a percentage. The number is multiplied by 100 and the % character is appended.
, (comma)	This is the thousands separator. If the format contains commas separated by #’s or 0’s, the number is displayed with commas separating thousands. A comma following a placeholder scales the number by a thousand. For example, the format 0, scales the number by 1000 (e.g., 10,000 would be displayed as 10).
E- E+ e- e+	Displays the number in scientific notation. If the format contains a scientific notation symbol to the left of a 0 or # placeholder, the number is displayed in scientific notation and an E or an e is added. The number of 0 and # placeholders to the right of the decimal determines the number of digits in the exponent. E- and e- place a minus sign by negative exponents. E+ and e+ place a minus sign by negative exponents and a plus sign by positive exponents.
$-+/(): space	Displays that character. To display a character other than those listed, precede the character with a back slash “\” or enclose the character in double quotation marks. You can also use the slash “/” for fraction formats.
\	Display the next character. The backslash is not displayed. You can also display a character or string of characters by surrounding the characters with double quotation marks. The backslash is inserted automatically for the following characters: ! ^ & ‘ ’ `~ { } = < >
* (asterisk)	Repeats the next character until the width of the column is filled. You cannot have more than one asterisk in a format section.

Format Symbol	Description
_ (underline)	Skips the width of the next character. For example, to make negative numbers surrounded by parentheses align with positive numbers, you can include the format _) for positive numbers to skip the width of a parenthesis.
“text”	Displays the text inside the quotation marks.
@	Text placeholder. If there is text in the cell, the text replaces the @.
m	Month number. Displays the month as digits without leading zeros (e.g., 1-12). Can also represent minutes when used with h or hh formats.
mm	Month number. Displays the month as digits with leading zeros (e.g., 01-12). Can also represent minutes when used with the h or hh formats.
mmm	Month abbreviation. Displays the month as an abbreviation (e.g., Jan-Dec).
mmmm	Month name. Displays the month as a full name (e.g., January-December).
d	Day number. Displays the day as digits with no leading zero (e.g., 1-31).
dd	Day number. Displays the day as digits with leading zeros (e.g., 01-31).
ddd	Day number. Displays the day as an abbreviation (e.g., Sun-Sat).
dddd	Day number. Displays the day as a full name (e.g., Sunday-Saturday).
yy	Year Number. Displays the year as a two-digit number (e.g., 00-99).
yyyy	Year Number. Displays the year as a four-digit number (e.g., 1900-2078)
h	Hour number. Displays the hour as a number without leading zeros (e.g., 1-23). If the format contains one of the AM or PM formats, the hours are based on a 12-hour clock. Otherwise, they are based on a 24-hour clock.
hh	Hour number. Displays the hour as a number with leading zeros (e.g., 01-23). If the format contains one of the AM or PM formats, the hours are based on a 12-hour clock. Otherwise, they are based on a 24-hour clock.
m	Minute number. Displays the minute as a number without leading zeros (e.g., 0-59).
mm	Minute number. Displays the minute as a number with leading zeros (e.g., 00-59). The mm format must appear immediately after the h or hh symbol. Otherwise, it is interpreted as a month number.
s	Second number. Displays the second as a number without leading zeros (e.g., 0-59).
ss	Second number. Displays the second as a number with leading zeros (e.g., 00-59).
AM/PM, am/pm, A/P, a/p	12-hour time. Displays time using a 12-hour clock. Displays AM, am, A, or a for times between midnight and noon; displays PM, pm, P, or p for times from noon until midnight.
[*color*]	Displays the output in the specified color where *color* is BLACK, BLUE, CYAN, GREEN, MAGENTA, RED, WHITE, or YELLOW.

Format Symbol	Description
[COLOR n]	Displays the text using the corresponding color in the color palette. n is a color in the color palette.
[conditional value]	Under normal circumstances, the four sections of a custom format are for positive numbers, negative numbers, zeros, and text. Use brackets to indicate a different condition for each section. For example, you might want numbers less than three to be black and those greater than three to be red. You would use: [>3] [RED][General] ; [<-3][RED][General];[BLACK][General]

The following table shows some examples of how these formats are displayed.

Format	Cell Data	Display
#.##	654.429	654.43
#.##	0.429	.43
#.0#	654	654.0
0.00	654.429	654.43
0.00	654.4	654.40
0.00	.429	0.43
#,##0“CR”;#,##0“DR”;0	654.429	654CR
#,##0“CR”;#,##0“DR”;0	-654.429	654DR
#,	10000	10
“Resid=”0.0	123.45	Resid=123.5
000-00-0000	123456789	123-45-6789
m-d-yy	2/3/94	2-3-94
mm dd yy	2/3/94	02 03 94
mmm d, yy	2/3/94	Feb 3, 94
mmm d, yyyy	2/3/94	Feb 3, 1994

Data Type

The sixth column of the Variable Info sheet gives the variable's data type. Currently, there are five data types:

0. Text - case used. (default with Data Type left blank)
1. Text - case ignored
2. Numeric
3. Month (sorts alphabetic month values in January through December)
4. Fixed (not rearranged by sorting or insert/delete; usually used for value label information)

The Data Type of a variable is only used when the variable is employed as a grouping or break variable, as in cross tabulation. In these cases, the data type specifies how the group values are sorted. For example, suppose a variable contains the numeric values: 11, 6, 10, 1, 22, 7. When treated as text, these numbers are sorted in the order: 1, 10, 11, 22, 6, 7. If you designated the Data Type as Numeric, these values would be sorted in the usual numeric order: 1, 6, 7, 10, 11, 22.

The Data Type is ignored when the variable is used in numeric calculations.

Value Label

The seventh column of the Variable Info sheet specifies the value labels. Value labels are labels that are displayed in the place of the original value on the database. For example, you might be tabulating a survey in which a Yes was coded as a 1 and a No was coded as a 0. The final report will be much more interpretable if the Yes and No are displayed rather than the 1 and 0. One approach is to generate a new variable using a Find/Replace operation or the Recode transformation to replace the 1's and 0's with Yes's and No's. A simpler approach is to use Value Labels.

Creating Value Labels

Two methods are available for creating value labels. The original method was to store a list of possible values and their labels directly on the spreadsheet. A new method, which was recently added, is to store the values and their labels in a text file. We recommend using the second method. We will now explain each approach.

Creating Value Label Files (Recommended)

Values label files are easily constructed using Windows Notepad or similar text editor. The files contain a list of values and their corresponding labels, one set per row. The value and the label must be separated by a tab (not a blank since the labels may contain blanks). Note that this is the format of data that is copied and pasted into a text file from a spreadsheet such as **Excel** or ***GESS***.

Here is an example of a value label file called Likert.txt:

```
1	Strongly Disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly Agree
```

Attaching the Value-Label File to Variables

Once you have created a value-label file, you must associate it with those variables whose values you want to label. This is accomplished by entering the name of value label file (Likert.txt in this example) on the *Variable Info* sheet in the *Value Label* column. Note that several variables can share the same value-label file.

Step-by-Step Instructions for Using Value Labels

1. Click on the *Variable Info* tab at the bottom of the spreadsheet.
2. Double-click in the *Value Label* column on the row corresponding to the variable you want to label.
3. Click the button *Create/Edit a V.L. File* to run the Notepad program.
4. Enter the values and their labels, one per line. The syntax is *Value <tab> Label*. The order of the values does not matter. Note that intermediate blanks in the values and the labels are retained and treated like any other characters.
5. After completing the value label file, save it into the same directory in which the database resides. Note that ***GESS*** first searches for a value label file in the directory in which the database resides and then in the ***GESS*** Data directory. Next, exit the Notepad program. We suggest that you use '.txt' as the extension for value label files. This will make them easy to find.
6. Press the *Select a V.L. File* button, select the appropriate file, and press the *Open* button to complete the selection.
7. Back on the *Specify the Value Labels* window, click the *Ok* button to complete the selection.
8. This completes the specification of the value label file for this variable. Note that if you knew the name of a previously created file, you could have just typed it in directly into the space provided.
9. To use the value labels in a specific procedure, you must set the Value Labels option (usually under the Format or Reports tabs) to *Value Labels*.

Creating Value Label Variable in the Database (Not Recommended)

Values labels are constructed from two contiguous variables. The variable on the left contains the original value. The variable on the right contains the corresponding label. For example, you might have several variables whose values range from 1 to 5. Suppose you want to label these values as shown below. Further suppose that variables C11 and C12 are empty columns on the spreadsheet. You would enter the five possible values in variable C11 and the corresponding labels in C12, as shown below.

C11	C12
1	Strongly Disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly Agree

Attaching the Value-Label Variables to Other Variables

Next, you must associate this set of value labels with those variables that contain this type of data. This is accomplished by entering the value label variable (C11 in this example) on the *Variable Info* sheet in the *Value Label* column. For example, suppose you wanted to attach this set of value labels to variable C4. You would specify C11 in the Value Label column of the Variable Info sheet in the row corresponding to C4. Now, when you use C4 in reports that display individual values (like crosstabs), the value labels in C12 will be displayed.

Note that the values of the original variable need not be numeric. You can associate value labels with a text values as well numeric. Also note that several variables can use the same value labels.

Copying Value Labels among Databases

Value labels variables must be included on each database. You cannot read the value labels from another database. If you have a set of value label variables that you wish to use on more than one database, you will have to Copy and Paste them from one database to another. This is done by copying the value-label variables to the clipboard, opening the second database, and pasting the clipboard information onto it.

Rows / Observations

The rows of the database correspond to the observations or cases. Normally, you begin adding data at the top of the datasheet and work your way down.

Currently, there is a limit of 16,382 rows per S0 database. S0Z databases may have as many rows as your hard disk will hold.

Chapter 102

Spreadsheets

Introduction

This chapter discusses the operation of the *Spreadsheet*, one of the three main windows of the **GESS** system. The other two windows, the Output window and Procedure window will be discussed in other chapters. The Spreadsheet is the window that lets you enter, view, and modify your data. It is the first screen that you are presented with when you start the program.

The operation of the **GESS** spreadsheet is similar to the operation of other spreadsheets with which you are familiar. In fact, it has most of the operational features of Microsoft Excel. Since the operation of these spreadsheets is so common, we will not spend a lot of space teaching them to you. We will now go over the details.

First, we will discuss the menu bar which appears at the top of the spreadsheet window.

Spreadsheet Menus

You should be familiar with the operation of pulldown menus. We will discuss the various options that are on these menus.

File Menu

The File Menu controls the opening and closing of databases. Note that some of the basic File Menu operations are also provided on the Toolbar.

We will now discuss each of these options.

- **New**

  This option closes the current database (if any) and creates a new database. A dialog box appears that lets you select the type of database: S0 or S0Z. See the *Introduction* to the *Database* chapter for details on the differences between these two types of databases.

  When you create a new S0Z database, an expanded dialog box will appear that will allow you to specify the name of the database, the number of variables (columns), and the maximum length of the numeric and text values. You can increase the number of variables and the length of the record using the *Insert-Columns* operation.

- **Open**

  The Open option lets you open existing ***GESS*** databases. It will cause the Open Dialog box to appear from which you can select a file. If you want to open a S0Z database, change the type of files that are scanned for by the dialog box. These files will have the extension “S0Z.”

  When selecting an S0 database, note that you select the file with the *S0* extension. Attempting to open ***GESS*** files with other extensions (such as *S1*, *S2*, etc.) will produce unpredictable results.

Note that the database is copied into memory. Once you open a database, you actually have two copies of it—one in memory and one on your disk. No automatic relationship is maintained between the loaded database and the disk database. Changes made to the copy in memory will not automatically change your disk database files unless you save them!

- **Add a Sheet**

  Each datasheet contains 256 variables. This option lets you add an additional datasheet to your database. When selected, an additional Datasheet Tab will appear at the bottom of the spreadsheet and additional variables are added to the Variable Info sheet.

- **Import**

  This option lets you import data from various other spreadsheets, databases, and statistical systems into ***GESS***. There is a whole chapter devoted to this topic, so refer to that chapter for further details.

- **Remove Last Sheet**

  When your database (spreadsheet-type only) contains two or more datasheets, this option will remove the last one from the database. Note that the datasheet must be blank before it can be removed. If the last datasheet contains data, select the data and cut it to remove it. This will clear the datasheet so that it can be removed.

- **Printer Setup**

  This option brings up a window that lets you set parameters of your printer(s).

- **Page Setup**

  This option lets you specify the format of your datasheet printout. You can specify headers, footers, margins, page order (across or down), and scale (size of the print). It is often used to enable the printing of the row and column labels.

  Headers and footers can contain text and special formatting codes. The following table lists the special formatting codes. Header and footer codes can be entered in upper or lower case.

Format Code	Description
&L	Left-aligns the characters that follow.
&C	Centers the characters that follow.
&R	Right-aligns the characters that follow.
&D	Prints the current date.
&T	Printers the current time.
&F	Prints the worksheet name (this is an internal name that may not be useful).
&P	Prints the page number.
&P+*number*	Prints the page number plus number.
&P-*number*	Prints the page number minus number.
&&	Prints an ampersand.
&N	Prints the total number of pages in the document.

The following font codes must appear before other codes and text or they are ignored. The alignment codes (e.g., &L, &C, and &R) restart each section; new font codes can be specified after an alignment code.

Format Code	Description
&B	Use bold font.
&I	Use an italic font.
&U	Underline the header.
&S	Strikeout the header.
&”fontname”	Use the specified font.
&nn	Use the specified font size - must be a two digit number.

- **Print**

  This option prints the entire datasheet or a portion that you designate. Note that you must print each datasheet of a database separately. If the row and column labels do not show in your printout, select Page Setup from the File menu and check the appropriate selection boxes.

- **Save**

  This option saves the current database. Remember that a database consists of several files. All of those files will be replaced.

- **Save As**

  This option saves the current database to a database with a different name. For example, if you are working on a database called “XYZ” and will be making changes to it, you might want to save a copy of it as “XYZ1” so that any mistakes you might have made will not destroy your original data.

- **Export**

  This option lets you export the current database to another format for use in other spreadsheets, databases, etc. There is a chapter devoted to this topic, so we refer you to that chapter for further details.

- **Exit GESS**

  This option quits the ***GESS*** system.

- **Previously Open Files**

  A list of previously open databases is presented. You may select any of them to revert to that database directly.

Edit Menu

The Edit Menu controls the editing of databases. Note that some of the basic Edit Menu operations are provided on the Toolbar.

- **Undo**

  Undo allows you to undo the last edit operation made. Note that only the most recent edit operation may be undone. Also not that the undo only works for S0 databases. It does not work for S0Z databases.

  When you make wholesale changes to your database (by cutting or pasting, for example), the Undo system requires a lot of memory to store additional information needed for an undo operation. If you see system resources getting low, make an additional change to a single cell. This will reset the undo system and free up system memory.

- **Cut**

  The Cut option copies the currently selected (highlighted) data to the Windows clipboard and clears those cells. This data may be pasted at another location within ***GESS*** or to another Windows program. We will discuss the process of selecting cells later in this chapter.

- **Copy**

  The Copy option copies the currently selected (highlighted) data to the Windows clipboard. The selected data is untouched. The copied data may be pasted at another location within ***GESS*** or to another Windows program. We will discuss the process of selecting cells later in this chapter.

- **Paste**

  The Paste option copies data from the clipboard to the current datasheet at the currently selected location. The contents of the clipboard may have come from a previous Cut or Copy operation within ***GESS*** or from another Windows program.

  For example, an easy way to analyze the means from an analysis of variance is to copy them from the Output screen and paste them in a datasheet. Furthermore, a quick way to import data from an *Excel* spreadsheet is to copy the data in *Excel* and paste it into an **GESS** datasheet.

  The Paste option behaves differently depending on whether part of the datasheet is selected (highlighted). The data on the clipboard acts like a rectangular array of data (it has rows and columns). If there is no selection on the datasheet, the paste operation will place the data on your spreadsheet just as it appeared when it was copied to the clipboard. However, if the spreadsheet has a selected area, the paste operation will do two things. First, it will insert the data inside the selected area only (extra data will be omitted). Second, it will repeat either all or part of the clipboard so that the selected area is filled.

- **Paste Rotated**

  This option works like the Paste option (above), except that the data are rotated ninety degrees so that the rows become the columns and the columns become the rows. The following example shows the result of using this option:

  Data that was Copied

1	2
3	4
5	6
7	8

  Result of Paste Rotated Operation

1	3	5	7
2	4	6	8

- **Clear**

  The Clear option erases the data that is selected. Note that unlike the Cut option, the Clear option does not put the data on the clipboard.

- **Insert**

  The Insert option inserts rows or columns into your datasheet at the current position of the cursor. When you select Insert, a dialog box appears that allows you to indicate whether you want to insert rows or columns (variables). The number of rows (or columns) inserted is determined by the number of rows (or columns) selected.

  Hence, the steps to insert columns are as follows:

  1. Select the number of columns you want to insert, beginning your selection at the column where you want them added.
  2. Select Insert from the Edit Menu.
  3. Select Columns from the dialog box.

  A datasheet on an S0 database contains exactly 256 variables. When you insert new variables, the current variables are shifted to the right. The variables at the right of the datasheet are "pushed off" the datasheet and lost. For this reason, ***GESS*** first checks to make sure that there are enough empty variables at the edge of the datasheet to accommodate the inserted variables. If you find that you don't have room to insert variables that you need, simply add a new datasheet, cut the last several variables at the right of the datasheet, and paste them to the next datasheet. This will make room for inserting variables.

  When you insert columns in a S0Z database, a new database is created with the new column count. The record size is increased by ten bytes for each new variables (assuming that you are using the default ten bytes per number).

- **Delete**

  The Delete option removes the currently selected rows or columns from your datasheet. When you select Delete, a dialog box appears that lets you indicate whether to delete rows (or columns). The number of rows (or columns) deleted is determined by the number of rows (columns) selected.

- **Fill**

  This option brings up the Fill window which fills the current variable (or the currently selected block of cells) with the value specified. The value may be incremented so that special patterns such as 1,2,3,4 may be easily generated.

- **Find**

  This option searches through your data for a designated value. Once you have started a find operation, use the Find Next button continues your search. You can search for a single digit or for the complete number.

- **Replace**

  The Replace option allows you to quickly replace data throughout your datasheet. You can replace only those cells that match a certain pattern, or you can replace individual letters and digits.

- **Variable Info**

  This main brings up a submenu that lets you modify information about the variable such as its name, label, format, data type, or transformation. Each of the items on the submenu call up dialog windows that let you modify the corresponding information.

- **Font**

  This option allows you to change the font of cells in the database.

- **Options**

  This brings up a window that allows you to change various options that control the way the program functions.

- **Serial Numbers**

  This brings up a window displaying the serial numbers being used and the product(s) licensed.

- **Enter File Name F7**

  This option launches the Windows dialog used to find a file on your computer. The file name is entered into the active cell on the spreadsheet. This is commonly used when entering file names in ***GESS*** procedures.

Data Menu

The Data Menu lets you display or modify the database. It also lets you reduce the number of rows that are processed using the Filter procedure.

- **Data Report**

  This option brings up the Data Report procedure which lets you create a nicely formatted printout of all or part of your data.

- **Sort**

  The Sort option lets you sort a datasheet by up to three variables. The sort may be either ascending or descending on a sorted variable. To sort your data, simply select the variables you want to sort on and click the Ok button.

Note that the sort procedure rearranges only the datasheet on which the variables reside. Other sheets are untouched.

- **Filter**

  This section explains how to use *filters* to limit which rows (observations) are used by the other procedures and which are skipped. For example, you might want to limit an analysis to those weighing over 200 pounds. You would use a filter to accomplish this.

  The filter tab is used to enter the desired filter statements, as well as all filter options.

Filter Specification

Filter System Active

This statement must be checked for the Filter to be activated.

Note: You must RUN this screen to activate (or de-activate) the Filter System.

Keep Spreadsheet Row If:

You can specify several filter expressions. This option specifies how these expressions are combined.

- **OR**

  If you select the 'OR' option, the condition on only one of the expressions must be met to retain the row in the analysis.

- **AND**

  If you select the 'And' option, the conditions in all of the expressions must be met in order to retain the row in the analysis.

Filter Statements

These boxes contain the filter statements. Each box may contain several filter expressions separated by semicolons.

Note that text must be enclosed in double quotes.

Example: C1<5; C2=4; C3=1,2,3; C4<C5; C6<C7+C8; C1<>Missing

SYNTAX:

The basic syntax of a filter statement is

VARIABLE LOGIC OPERATOR VALUE

where VALUE is an expression the yields a text value or a number constructed from variable names, numbers, and the symbols +, -, *, and / (add, subtract, multiply, and divide). Note that parentheses are not allowed. A list of values may be used.

Examples of valid VALUE expressions are

C1

Height

1.0

C1+5

Height/100

1,2,3,4,5

X+Y/100+4

Missing (may be used to indicate a missing value)

“John” (must be enclosed in DOUBLE quotes)

LOGIC OPERATOR

The Logic Operator is one of the following operators:

=	Equal to
<>	Not equal
<	Less than
<=	Less than or equal
>	Greater than
>=	Greater than or equal

Note that only one operator may be specified in an expression.

Variable Name Locator (for pasting variable names into filter statements)

The selected variable name may be copied to the clipboard and pasted into the filter statement.

This field provides no active options. It is here to let you have easy access to a list of all variable names on the current database.

Filter Specification - Filter Statement Comparison Option

Comparison Fuzz Factor

When you make a comparison, you may want to allow for a certain amount difference between two numbers that may occur because of rounding error, etc. For example, you may want the statement .3333=.3334 to evaluate to true instead of false. If the fuzz factor is set to .000001, this expression will be false. However, if the fuzz factor is set to .0001, then this expression will be true.

- **Merge Databases**

  This option brings Merging Two Databases procedure.

- **Enter Transform**

  This option brings up the Transformation window. A discussion of transformations is the subject of another chapter.

- **Recalc Current**

  The variable at which the cursor is located is recalculated using the transformation formula stored in the Variable Info sheet. Data that are currently in this column are replaced with the transformed values.

- **Recalc All**

  The whole database is recalculated, proceeding from left to right. Note that data that are currently stored in the database are replaced by the transformed values. If a variable does not have a transformation formula, its values are untouched.

- **If-Then Transform**

  This brings up the If-Then Transformations window. A discussion of if-then transformations is the subject of a later chapter.

- **Data Simulation**

  This option brings up the Data Simulator.

Analysis and Graphics Menus

These menus load the corresponding procedure windows. For example, select Descriptive Statistics, then Data Screening will load the Data Screening procedure window. This window controls the running of the Data Screening procedure.

Tools Menu

From this menu you can load the Macro Command Center, the Data Simulator window, or the Merging Two Databases window.

Window Menu

This menu lets you transfer to one of the other **GESS** menus such as the Output window or one of the currently open procedure windows.

Help Menu

From this menu you can launch the **GESS** Help System and view PDF documentation, tutorials, and references. From this menu you can also view serial numbers and licensing information.

Spreadsheet Toolbar

The toolbar is provided for single-click access to the most commonly used menu options. You will find that each of the options on the toolbar can also be found in the menus. The toolbar has a feature called a "tool tip." This means that when you hold the mouse pointer over a certain square for at least a second, a small help box will appear that explains what this particular toolbar button is for. Most of the buttons on the toolbar follow Windows standards, so you will recognize them right away.

The last eight buttons on the toolbar represent procedures that you may transfer to. These are completely customizable. You can designate which eight procedures you want to be able to transfer to by right clicking on them in the Navigator window.

Cell Reference

Two boxes at the bottom of the screen give the *Cell Reference*. The first box gives the variable (column) number of the current location of the cell cursor. The second box gives the row number of the current location of the cell cursor. The cell cursor is the active cell. You can recognize the current cell because it will have an extra dark border.

Cell Edit

The Cell Edit box provides an alternate place to edit data. As you move around the spreadsheet, the contents of the active cell are copied to this Cell Edit box. Occasionally, your data will be longer than can easily be displayed in the cell. Although you could reset the column width, you usually find it easier to edit the data in the Cell Edit box.

Any changes you make will not be entered into the datasheet until you hit the ENTER key or position the mouse on another cell. After making changes to data in the Cell Edit box, you can press the ESC key to withdraw the changes.

Datasheet

This section of the screen shows the data. We will now describe how to use the spreadsheet to modify the data contained in a datasheet. You should know how to select cells, ranges, rows, and columns. Your work will go much faster if you learn how to quickly enter, modify, and delete data. These will all be described in this section.

Navigating the Datasheet

This section describes how to move around the datasheet using the keyboard and the mouse. In addition to moving around the datasheet, we will also describe how to make selections, copy data, and move data.

Active Cell

The datasheet cursor is always located on a single cell, even when a range of cells is selected. The cell on which the datasheet cursor is located is called the *Active Cell*. Any typing that is done will only affect the active cell. The contents of the active cell are displayed in the Cell Edit box. The address of the active cell is displayed in the Cell Reference boxes.

Keyboard Commands

The following commands are used mainly for data entry.

Key	Description
ENTER	Accepts the current entry and moves the active cell down one cell. When a range of cells is selected, accepts the current entry and moves down to the next selected cell. When the bottom of the selection is reached, the active cell moves to the top of the next selected column to the right.
SHIFT-ENTER	Acts like the ENTER key, except that cell-to-cell movement is upward and to the left instead of downward and to the right.
TAB	Accepts the current data entry and moves the cursor one cell to the right. When a range of cells is selected, the tab moves the cursor to the right to the next cell in the selection.
SHIFT-TAB	Acts just like the TAB key, except that cell-to-cell movement is to the left instead of to the right.
F2	Enters edit mode. Pressing F2 a second time brings up a cell-text data entry box.
DEL	Clears the current entry or selection.
ESC	Cancels the current data entry.

The following commands are used mainly for moving about the datasheet.

Key	Description
UP ARROW	Moves the active cell up one row.
DOWN ARROW	Moves the active cell down one row.
LEFT ARROW	Moves the active cell left one column.
RIGHT ARROW	Moves the active cell right one column.
CTRL UP / DOWN / LEFT / RIGHT ARROW	Moves to the next range of cells containing data. If there is no additional data in any of the cells in that direction, the active cell is moved to the edge of the datasheet.
PAGE UP	Moves up one screen.
PAGE DOWN	Moves down one screen.
CTRL PAGE UP	Moves left one screen.
CTRL PAGE DOWN	Moves right one screen.
HOME	Moves to the first column in the current row.
END	Moves to the last column in the current row that contains data.
CTRL HOME	Moves to the upper-left corner of the datasheet (cell 1,1).
CTRL END	Moves to the last row and column that contains data.
SCROLL LOCK	Modifies the action of the above movement keys. This key causes the datasheet window to scroll without changing the current selection. It works with all movement keys except HOME, END, CTRL HOME, and CTRL END.
SHIFT + any movement key	Extends the current selection in the direction indicated.

Mouse Actions

The mouse is used mainly for positioning the active cell and making selections. You can also use the mouse to move a block of cells around the datasheet.

Action	Description
Left Click	Accepts the current entry and moves the active cell to the position of the mouse.
Right Click	Does nothing.
Left Click in a Row or Column Heading	Selects the entire row or column.
Left Double Click	In-cell editing is invoked.
Right Double Click	Does nothing.
Left Click and Drag	Selects a range of cells. If other ranges were selected, they are unselected.
CTRL + Left Click and Drag	Selects a range of cells. If other ranges were selected, they remain selected. Note that edit commands, such as cut and copy, will only work on a single range.
SHIFT + Left Click and Drag	Extends the current selection in the direction indicated.
Dragging a Selection’s Copy Handle	Copies the selection to a new location. The copy handle is the small plus sign at the lower-right corner of a selection.
Dragging a Selection’s Border	Moves the selection to a new location. Note that when you move data around the datasheet in this fashion, no attempt is made to update the variable names. You will have to do that manually. For this reason, it is best to do this kind of wholesale editing before you attach names and transformations to the variables. Note also that Cell Transformation formulas are updated.

Selecting Cells

Many operations require one or more cells to be selected. There are three types of selections: a single cell, a single rectangular range of cells, and multiple ranges of non-adjacent cells. Cells may be selected either with the mouse or with the keyboard.

Selecting Cells with the Mouse

To select a range of cells with the mouse, click and hold the left mouse button down on the upper-left cell of the range you want to select. Drag the mouse cursor to the lower-right cell, while continuing to hold the left mouse button down. When the desired cells are selected, release the mouse button.

To select multiple ranges with the mouse, press the CTRL key while making each additional selection. Note that multiple selections are only useful for controlling cell cursor movement during data entry. They are not used by any of the edit functions.

To select an entire row or column, click on the row or column heading.

Once a range is selected, you can move the active cell within the selection using the Enter, SHIFT+ENTER, TAB, and SHIFT+TAB keys without destroying the selection.

Selecting Cells with the Keyboard

To select a range of cells with the keyboard, position the active cell at the upper-left corner of your desired selection. While holding down the SHIFT key, use the cursor movement keys (such as the arrow keys) to move to the lower-right corner of the selection.

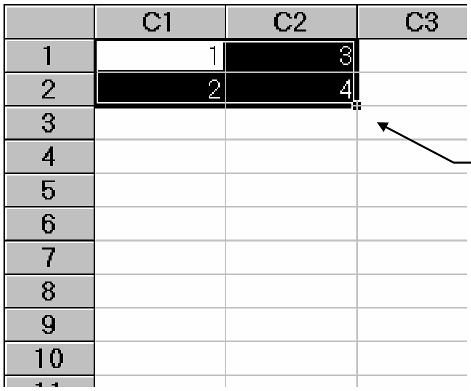
Editing Datasheets Interactively

Data can be entered into a datasheet in many different ways. In this section, we will explain how you can quickly move and copy ranges of cells by clicking and dragging the copy handle of a selection.

Copying Data Interactively

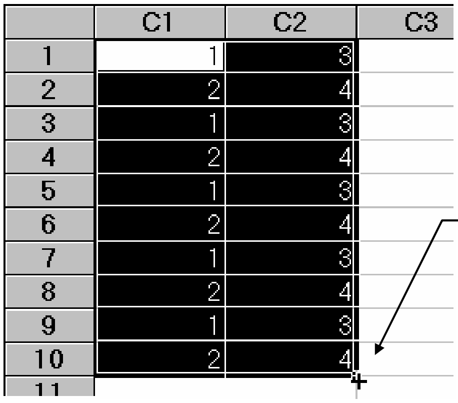
You can copy a range of cells quickly by using only a few clicks of your mouse. The steps for doing this are enumerated next followed by figures that illustrate the action.

1. Select a range of one or more cells.
2. Select the *copy handle* (the small crosshair that appears at the lower-right corner of a selection) with your mouse cursor by positioning the mouse cursor over the copy handle and pressing the left mouse button.
3. Drag the copy handle through the range of cells that are to receive the copied data.
4. Release the left mouse button.



	C1	C2	C3
1	1	3	
2	2	4	
3			
4			
5			
6			
7			
8			
9			
10			

The copy handle is at the lower right corner of the selection.



	C1	C2	C3
1	1	3	
2	2	4	
3	1	3	
4	2	4	
5	1	3	
6	2	4	
7	1	3	
8	2	4	
9	1	3	
10	2	4	
11			

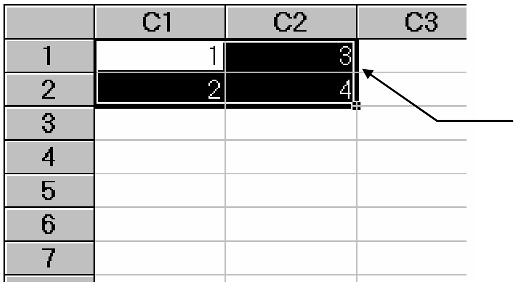
The cursor changes to a crosshair as the copy handle is dragged down.

Note that the copied selection is repeated so that the whole copied area has data.

Moving Data Interactively

You can move a range of cells quickly by using only a few clicks of your mouse. The steps for doing this are enumerated next followed by figures that illustrate the action.

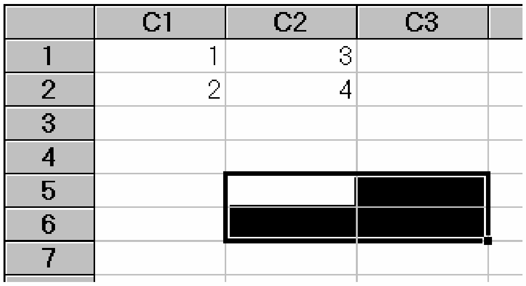
1. Select a range of one or more cells.
2. Position the mouse cursor on the border of the selection. When positioned on the border, the pointer changes to an arrow.
3. Drag the selection to the new location.
4. Release the left mouse button.



	C1	C2	C3
1	1	3	
2	2	4	
3			
4			
5			
6			
7			

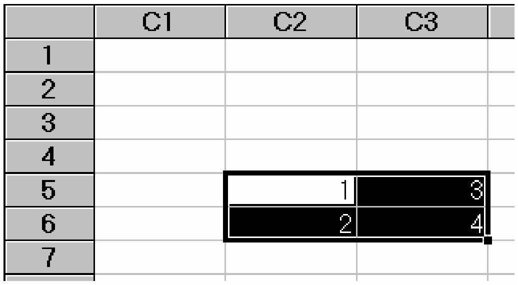
First, select the cells you wish to move.

Next, position the mouse pointer at the border of the selected area. The mouse pointer will change to an arrow. Once the pointer has changed, do not release the mouse button.



	C1	C2	C3	
1	1	3		
2	2	4		
3				
4				
5				
6				
7				

Move the selected cells to their new location.



	C1	C2	C3
1			
2			
3			
4			
5		1	3
6		2	4
7			

Release the left mouse button. The contents of the selected cells will move to the new location.

Caution: the variable names have not changed. You will have to manually adjust variable names and transformations when you move data using this method.

Changing Row Heights and Column Widths

You can interactively resize the height of a row or the width of a column using the mouse. Position the pointer on the right edge of a column heading or the bottom edge of a row heading. The pointer will change shape. Simply drag the pointer to resize the row or column.

If multiple rows are selected when you resize a row, all selected rows are resized as you drag a row border. Multiple columns can be resized in like manner.

You can also set the size of a selected group of columns or rows to equal the size of another row or column. First, select the group of columns or rows you want to resize, including the column or row whose size you want to match. Next, click the right border of the column header or the bottom border of the row whose size you want to match. The columns or rows will all be resized.

Datasheet Tabs

The Datasheet Tabs at the bottom of the spreadsheet let you move quickly from one datasheet to another. These tabs are especially useful for allowing you to move to the Variable Info sheet to make changes to variable names, transformation, and formats.

Chapter 103

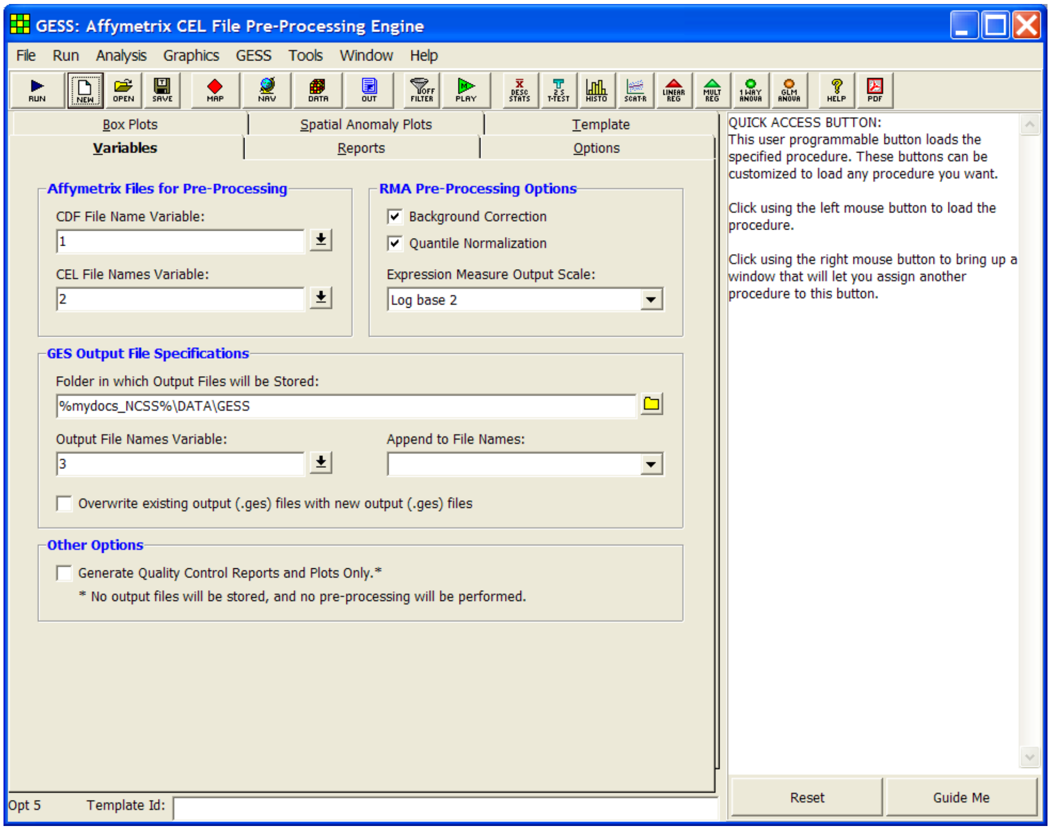
Procedures

Introduction

Each procedure has its own Procedure window. The Procedure window contains all the settings, options, and parameters that control a particular procedure. These options are separated into groups called *panels*. A particular panel is viewed by pressing the corresponding *panel tab* that appears just below the toolbar near the top of the window.

The current values of all options available for a procedure are referred to as a *template*. The template may be stored for future use in a *template file*. By creating and saving template files (often referred to as *templates*), you can tailor each procedure to your own specific needs. For example, the multiple regression procedure has about twenty different reports available. You can select the four or five reports that are useful to you and disable the rest. Your selection can be saved as a template. Each time you use the multiple regression procedure, you simply load your template and run the analysis. You do not have to set all the options every time. Templates are stored in the *C:\...\[My] Documents\NCSS\NCSS 2007\Settings* directory.

Note that up to six Procedure windows can be opened at a time.



Affymetrix Files for Pre-Processing

CDF File Name Variable: 1

CEL File Names Variable: 2

RMA Pre-Processing Options

☑ Background Correction

☑ Quantile Normalization

Expression Measure Output Scale: Log base 2

GES Output File Specifications

Folder in which Output Files will be Stored: %mydocs_NCSS%\DATA\GESS

Output File Names Variable: 3

Append to File Names:

□ Overwrite existing output (.ges) files with new output (.ges) files

Other Options

□ Generate Quality Control Reports and Plots Only.*

* No output files will be stored, and no pre-processing will be performed.

QUICK ACCESS BUTTON:
This user programmable button loads the specified procedure. These buttons can be customized to load any procedure you want.

Click using the left mouse button to load the procedure.

Click using the right mouse button to bring up a window that will let you assign another procedure to this button.

Opt 5 Template Id:

Default Template

Whenever you close a procedure, the current settings are automatically saved in a default template file named Default. This template file is automatically loaded when the procedure is next opened. This allows you to continue using the template without resetting all of the options.

Menus

File Menu

The File Menu is used for initializing, loading, and saving a copy of a template. Each set of options for a procedure, called a template, may be saved for future use. In this way, you do not have to set the options every time you use a procedure. Instead, you set the options the first time, save them as a template, and re-use the template whenever you re-use the procedure.

- **New Template (Reset)**

  This menu item resets all options to their default values.

- **Open Template Panel**

  This option sets the Template panel as the active procedure panel. This panel lets you load or save template files. It displays all templates associated with this procedure along with the Template Id (the optional phrase at the bottom of the window).

  Saving a Template

  To save a template, enter the name you want to give the template file in the File Name box. You may also enter an identifying phrase in the box at the bottom of the window since this will be displayed along side of the file names. Finally, press the Save Template button to save the file.

  Note that there is no automatic connection between the template in memory and the copy on the disk. If you want to save the changes you have made to a template, you must use the Save Template option to save them.

  Loading a Template

  To load a template file, select it from the list of files given in the Template Files box. Press the Load Template button to load the template.

- **Close Procedure**

  This option closes this Procedure window.

- **Save Template**

  This option saves the current option settings to the template file that is currently specified on the Template panel. You can be viewing any panel of the procedure when you issue this command—you do not have to be viewing the Template panel.

  You should note that the templates for each procedure have different file name extensions. Thus, you can use the same name for a template in the t-test procedure as for a template used in the multiple regression procedure.

Also note that the templates are stored in a subdirectory of the **NCSS** directory. You can erase any of the templates you want by deleting them from this directory.

Note the Save button on the toolbar provides this same operation. It may be more convenient than selecting this menu item.

- **Printer Setup**

  This option allows you to change the printer settings.

- **Exit**

  This option terminates the ***GESS*** system. Before using this option, you should save all datasheets and output documents.

Run Menu

The Run Procedure option runs the analysis, displaying the output in the Output document of the word processor. After you have set all options to their appropriate values, select this option to perform the analysis.

Note that the procedure may alternatively be run by pressing the F9 function key or by pressing the left-most key on the toolbar.

Analysis and Graphics Menus

These menus let you load any of the analysis or graphics procedure windows.

Tools Menu

From this menu you can load the Macro Command Center, the Data Simulator window, or the Merging Two Databases window.

Window Menu

This menu lets you display any of the other windows in the ***GESS*** system that are currently open such as the Output window, the Data window, the Navigator window, or any procedure window.

Help Menu

From this menu you can launch the ***GESS*** Help System and view PDF documentation, tutorials, and references. From this menu you can also view serial numbers and licensing information.

Entering Options

Your settings and selections are entered on the panel. The panel consists of several types of windows objects such as text boxes, check boxes, list boxes, and buttons. Each of these is used in the normal fashion.

Entering Text

When text (either numeric or letters) is needed for a particular option, you will be allowed to type text in the box. Many of these text boxes also have a pull-down button on the right. Pressing this button will allow you to select an option from a list of typical values, rather than type in the value.

Selecting from a List

Some options require you to select from a list. In this case, a dropdown list will allow you to choose from the selections available.

Selecting One or More Variables

When an option needs one or more variable names, you can type the names directly into the box or you can double-click in the box to bring up the Variable Selection window. This window lets you select one or more variables from those on the current database.

When using the Variable Selection window, you select one or more variables from the top window and they are listed in the button window. You can use the Shift key to select a list of contiguous variables. Use the Ctrl key to select disjoint (non-contiguous) variables.

At times, it may be more convenient to store the variable numbers rather than the variable names. Use the Store as Number button to indicate how you want the variables stored.

Toolbar

The toolbar is a series of small buttons that appear just below the menus at the top of the Procedure window. Each of these buttons provides quick access to a menu item. For example, the first button performs the same action as selecting the Run Procedure item from the Run menu.

Near the end of these buttons is a series of eight buttons that you can customize to represent your favorite procedures. This customization is done in the Navigator window. Pressing any of these buttons will load the corresponding Procedure window.

Chapter 104

Output

Introduction

GESS sends all statistics and graphics output to its built-in word processor from where they can be viewed, edited, printed, or saved. Reports and graphs are saved in rich text format (RTF). Since RTF is a standard document transfer format, these files may be loaded directly into your word processor for further processing. You can also cut and paste data onto a ***GESS*** datasheet for further analysis. This chapter covers the basics of our built-in word processor.

Documents

The ***GESS*** word processor maintains two documents: *Output* and *Log*. Although both of these documents allow you to view your data, the *Output* document serves as a viewer while the *Log* document serves as a recorder.

You can load additional documents as well. For example, you might want to view the output from a previous analysis to compare the results with the current analysis. To do this, you open a third document that is actually the log file from a previous analysis.

All ***GESS*** documents are stored in the RTF format. This is a common format that is used by most word processors, including Word and WordPerfect. When you save a ***GESS*** report, you will be able to load that report directly into your own word processor. All text, formatting, and graphics will appear in your word processor ready for further editing. You can then save the document in your word processor's native format. In this way, you can easily transfer the output of a ***GESS*** procedure to almost any format you desire.

Output Document

The Output document displays the output report from the current analysis. Whenever you run an ***GESS*** procedure (like t-test or histogram), the resulting reports and graphs are displayed in the Output document. Each new run clears the existing Output document, so if you want to save a report, you must do so before running the next report.

The Output document provides four main functions: display, print, save to the Log document, and save as an RTF file.

Log Document

The Log document provides a place to store a permanent record of your analysis. Since the Output document is erased by each new analysis, you need a place to store your permanent work. The Log document serves this purpose. When you have a report or graph that you want to keep, copy it from the Output document to the Log document.

The Log document provides four main word processing functions: load, display and edit, print, and save. When you load a file into the Log document, you can add new output to it. In this way, you can record your work on a project in a single file, even though your work on that project is spread out over several days.

Output Menus

You should be familiar with the operation of pull-down menus. We will discuss the various options that are on these menus

File Menu

The File Menu is used for opening, saving, and printing ***GESS*** word processor files. All options apply to the currently active document (the document whose title bar is selected). We will now discuss each of the options on this menu.

- **New**

  This option opens an empty document. You might use this when you want to make notes about your analysis.

- **New Log**

  This option opens an empty log document. You might use this when you want to start a new project.

- **Open**

  This option opens an existing file. When this item is selected, the Open Report File dialog box appears. Note that no connection is maintained between a loaded file and its image on the disk. If you make changes to a file, you must save those changes to the disk.

- **Open Log**

  This option opens an existing log file. When this item is selected, the Open Report File dialog box appears. The requested file is loaded into the Log document. Note that no connection is maintained between a loaded file and its image on the disk. If you make changes to a file, you must save those changes to the disk.

  You might use this option when you want to continue using a certain file as the Log file.

- **Toggle Auto-Log**

  When this menu item is checked, the output is automatically added to the bottom of the log file. If it is not checked, you must manually add the output to the log file by selecting “Add Output to Log.” To change this item from off to on or on to off, select it from the menu.

- **Add Output to Log**

  Selecting this option automatically copies the contents of the Output document to the Log document. The Output document remains unchanged. This allows you to save the current output document for further use.

- **Save As**

  This option lets you save the contents of the currently active document to a designated file using the RTF format. Note that only the active document is saved. Also note that all file names should have the “RTF” extension so that other systems can recognize their format.

- **Printer Setup**

  This option brings up a window that lets you set parameters of your printer(s).

- **Print Preview**

  This option allows you to preview the document before printing.

- **Print**

  This option lets you print the entire document or a range of pages. When you select this option, a Print Dialog box will appear that lets you control which pages are printed.

- **Close Output Window**

  Clears and minimizes the document. Note that this option will clear the Output and Log documents, but it will not close them since these two documents must remain open.

- **Exit GESS**

  This option exits the **GESS** system. All documents and databases are closed.

Edit Menu

This menu contains options that let you edit a document.

- **Undo**

  This item reverses the last edit action. It is particularly useful for replacing something that was accidentally deleted.

- **Cut**

  This item copies the currently selected text to the Windows clipboard and erases it from the document. You can paste the information from the clipboard to a different location in the current document, into another document, into a datasheet in the spreadsheet, or into another application. The selected text is erased.

- **Copy**

  This item copies the currently selected text from the document to the Windows clipboard. You can paste this information from the clipboard to a different location in the current document, into another document, into a datasheet in the spreadsheet, or into another application. The selected text is not modified.

- **Paste**

  This item copies the contents of the clipboard to the current document at the insertion point. This command is especially useful for moving selected information from the Output document to the Log document.

- **Select All**

  This item selects the entire document. Although you can select a portion of the document using the mouse or a shift-arrow key, this is much faster if you want to select the entire document.

- **Toggle Page Break**

  Changes the status of the page break on the line at which the insertion point resides. If a page break exists (shown by a horizontal line), it is removed. If a page break does not currently exist at that point, one is added.

  Note that **GESS** does not repaginate your document for you. Once you make changes, it will be up to you to repaginate your document.

- **Find**

  This item opens the Search dialog box. You can specify text that you want to search for. This is especially useful when you are looking for a certain topic or data value in a large report.

- **Find Next**

  This item continues finding the text you entered in the Search Dialog box.

- **Replace**

  This item opens the Search and Replace dialog box. This allows you to quickly make repetitive changes. For example, you might want to change the name of one of the variables to a more useful name.

- **Goto Section**

  This item does not modify the document. Instead, it lets you reposition the insertion point to one of the major topics. When **GESS** runs a procedure, it stores the major report topics in this list box. You can quickly position the view to a desired topic using this screen.

View Menu

This menu lets you designate which editing tools you want to use.

- **Ruler**

  This option controls whether the ruler and the tabs bar are displayed. The ruler displays the physical dimensions of the document. The tabs bar, found just below the ruler bar, lets you set the margins and tabs of your document. Only the currently selected part of your document is affected by a change in the tabs and margins.

- **Format Toolbar**

  The Format Toolbar lets you make formatting changes to the currently selected text. The function of each of the buttons is shown below.

- **Status Bar**

  The Status Bar shows the current position of the insertion point (cursor).

- **Show All**

  Selecting this menu item causes the ruler, tabs bar, format toolbar, and status bar to be displayed.

- **Hide All**

  Selecting this menu item causes the ruler, tabs bar, format toolbar, and status bar to be hidden. This gives you more screen space to view your output.

- **Redraw**

  Selecting this menu item causes the output to be redrawn.

Format Menu

This menu lets you set the format for a selected block of text.

- **Font**

  This option displays the Replace Font dialog box, which lets you specify the font and style of the selected text.

- **Paragraph**

  This option displays the Paragraph dialog box, which lets you specify the tabs and margins of the selected text.

- **Format Markers**

  Indicates whether the (usually hidden) tab arrows and the end-of-paragraph marks are displayed in the document. Note that these characters are never printed.

Window Menu

This menu lets you designate how you want the documents arranged on the screen and which ***GESS*** window you want displayed on top of your output desktop.

- **Cascade**

  This item arranges the documents in a cascading display from the upper left to the lower right of the screen.

- **Tile Horizontally**

  This item arranges the documents horizontally across the word processor window.

- **Tile Vertically**

  This item arranges the documents vertically down the word processor window.

- **Arrange Icons**

  When a document is minimized, it is represented as an icon at the bottom of the word processor window. This option arranges all document icons. It is usually applied when the word processor window has been resized.

- **Current Output**
  This item causes the Output window to be displayed on top of your desktop.

- **Log**
  This item causes the Log window to be displayed on top of your desktop.

- **View Data**
  Causes the Spreadsheet window to be displayed on top of your desktop.

- **Navigator**
  Causes the **GESS** Navigator window to be displayed on top of your desktop.

- **PASS Home**
  Causes the **PASS** Home window to be displayed on top of your desktop.

- **Quick Launch**
  Causes the Quick Launch window to be displayed on top of your desktop.

Help Menu

From this menu you can launch the ***GESS*** Help System and view PDF documentation, tutorials, and references. From this menu you can also view serial numbers and licensing information.

Chapter 105

Navigator and Quick Launch

Introduction

The **NCSS** Navigator window lets you quickly and easily find the appropriate statistical or graphical procedure. Designed in an outline format, it lists every procedure in the system along with a brief paragraph that describes what the procedure is for and when it might be used.

The Navigator window also lets you configure the eight procedure buttons that appear on the toolbars of the Data, Output, and Procedure windows. These buttons give you immediate access to your favorite procedures.

The **NCSS** Quick Launch window is an alternative method for finding statistical and graphical procedures. The Quick Launch window shows all the procedures of the program on a single tab. This window may also be used for configuring the eight procedure buttons of the toolbar.

Using the Navigator

The Navigator window is very easy to use. The Navigator window may be loaded by selecting the Navigator item from any Window menu or by clicking the globe icon on any of the toolbars.

The Navigator window has a set of menus, a toolbar, and then a large display area. On the left side of the display area is an outline list of all the statistical and graphical procedures in the system. On the right side of the display area is a window that will display a brief paragraph explaining the main purpose of the currently selected procedure.

Navigator Menus

Below is a description of each of the four menus that appear at the top of the Navigator window.

Outline Menu

- **Collapse Outline**
  This option collapses the outline so that only the main heading is displayed.

- **Expand to First Level**
  This option expands the outline so that the main headings and first-level subheadings are displayed.

- **Expand All**

  This option completely expands the outline so that all entries are displayed.

- **Bold Text**

  This option toggles the bolding of the text.

- **Goto Selected Procedure**

  This option loads the currently selected procedure's window.

- **Close the Navigator**

  This option closes the Navigator window.

Tools Menu

This menu allows you to open tool procedures such as Macros or Merge Databases.

Window Menu

This menu allows you to open other windows such as the Data window or the Output window.

Help Menu

This menu provides access to the help system, release date/version information, the serial number editing window, and printable electronic (PDF) documentation.

- **Help**

  This option launches the help system.

- **About**

  This option provides information about the release date and versions of ***GESS***, ***NCSS***, and ***PASS*** that you are currently using, the current filter status, and instructions for citing this software in publications.

- **View PDF**

  This option launches your PDF viewer to display the appropriate electronic documentation file.

Using the Quick Launch Window

The Quick Launch window may be loaded by selecting the Quick Launch item from any Window menu or by clicking the launch icon on any of the toolbars.

The Quick Launch contains a button corresponding to each statistical and graphical procedure in the system. As you mouse over each button, a brief paragraph explaining the main purpose of the currently selected procedure will appear in the message box to the right. The procedure name will also appear near the button.

A procedure window is launched by clicking on the corresponding button.

Documentation Access through Quick Launch

The Quick Launch window may be used to access the complete set of documentation in pdf format. Once the Quick Launch window is open, click on the Documentation tab. The pdf files are loaded by clicking on the corresponding pdf file button.

Toolbar

The toolbar gives you one-click access to several of the menu items. The menu item assigned to each button on the toolbar is displayed when the mouse is held over the button for a few seconds.

Customizing the Toolbars – Navigator

The eight procedure buttons that show up on all toolbars throughout the program may be changed from the Navigator. The process of assigning one of these eight buttons a new procedure is as follows:

1. Find and select the procedure in the outline section (left-side of main window) of the Navigator window.
2. Click on the button you want to assign the procedure to with the **right** mouse button.

The icon of the selected procedure will now appear in all toolbars throughout the program.

Customizing the Toolbars – Quick Launch

The eight procedure buttons that show up on all toolbars throughout the program may be changed from the Quick Launch window. To add (or change) a procedure to one of the eight toolbar buttons, click on the desire procedure, and drag it to the desired button on the Quick Launch toolbar. Release the mouse button. The new icon will replace the previous icon.

The icon of the selected procedure that was dragged and dropped will now appear in all toolbars throughout the program.

Chapter 106

Data Report

Introduction

This procedure generates a report of the data on a database. It is used when you want to maintain a printed copy of your data.

Data Structure

The procedure prints rows of selected variables. The rows printed may be selected using a filter.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables displayed on the report.

Data Variables

Data Variables

Select at least one variable to be printed. Both numeric and text data may be printed.

Report Options

Decimal Places

This option specifies the number of decimal places displayed for each variable. The number of decimal places is entered as a list of items separated by blanks or commas. For example, suppose you have selected variables X1, X2, and X3 for printing. If you enter “1,2,0” for this option, X1 will be printed with one decimal place, X2 with two decimal places, and X3 with no decimal places.

The number of decimal places can range from 0 to 9. In addition to this, you can enter one of three special formatting codes: **S**, **D**, and **F**.

S is used to indicate that numbers should be displayed in *single* precision.

D is used to indicate that numbers should be displayed in *double* precision.

F is used to indicate that numbers should be displayed using the *format* that is specified for the variable in the Variable Info sheet of the database. This allows you to specify commas, date conversions, etc.

For example, suppose you entered “F,S,2” here. X1 would be displayed using its format as specified on the Variable Info sheet, X2 would be displayed as a single precision number, and X3 would be displayed to two decimal places.

Note that if this statement is too short to account for all the variables, it is repeated. Hence, if you enter a “1” here and have three variables to display, all three will show a single decimal place.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want the table to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Precision

Specify the precision of numbers in the report. This is used when the format statement is left blank.

Label Justification

This option specifies whether the column labels should be right or left justified.

Data Justification

This option specifies whether the data should be right, left, or decimal justified.

Split Column Headings

Check this option to split the column headings into two headings instead of one.

Double Space

Check this option to add a blank row after each row.

Tabs

First

Specifies the position of the first item in inches. Note that the left-hand label always begins at 0.5 inches. Hence, the distance between this tab and 0.5 is the width provided for the row information.

Maximum

Specifies the right border of the report. The number of tabs is determined based on the First Tab, the Tab Increment, and this option. If you set this value too large, your table may not be printed correctly.

Increment

Specifies the width of an item in inches.

Offset

The labels are left justified. The data in the report are decimal tabbed (centered at the decimal place). Using two tabbing styles will cause the labels to be out of alignment with the data. Each data tab is moved to the right by this amount so that the data will line up with the column labels.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a List of Variables from a Database

This section presents an example of how to create a list of variables from a database. Data from the RESALE database will be used to generate the sample report.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Data Report window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **RESALE.s0**.
- Click **Open**.

2 Open the Data Report window.

- On the menus of the NCSS Data window, select **Data**, then **Data Report**. The Data Report procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

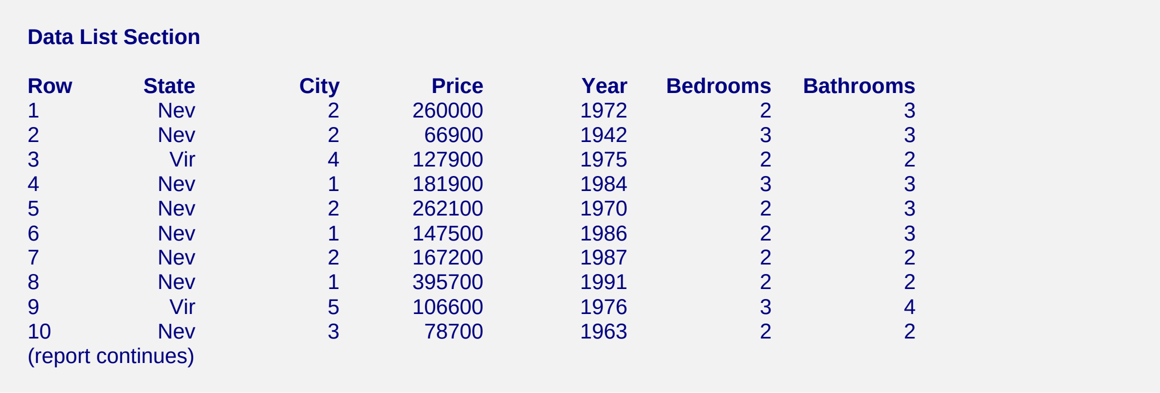
3 Specify the variables.

- On the Data Report window, select the **Variables tab**.
- Double-click in the **Data Variables** box. This will bring up the variable selection window.
- Select **State, City, Price, Year, Bedrooms,** and **Bathrooms** from the list of variables and then click **Ok**.
- Enter **0 0 0 0 0 0 0** in the **Decimal Places** box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data List Section



```
Data List Section

Row        State          City      Price        Year   Bedrooms   Bathrooms
1            Nev             2     260000        1972          2           3
2            Nev             2      66900        1942          3           3
3            Vir             4     127900        1975          2           2
4            Nev             1     181900        1984          3           3
5            Nev             2     262100        1970          2           3
6            Nev             1     147500        1986          2           3
7            Nev             2     167200        1987          2           2
8            Nev             1     395700        1991          2           2
9            Vir             5     106600        1976          3           4
10           Nev             3      78700        1963          2           2
(report continues)
```

This report lists the data in the selected variables.

Chapter 107

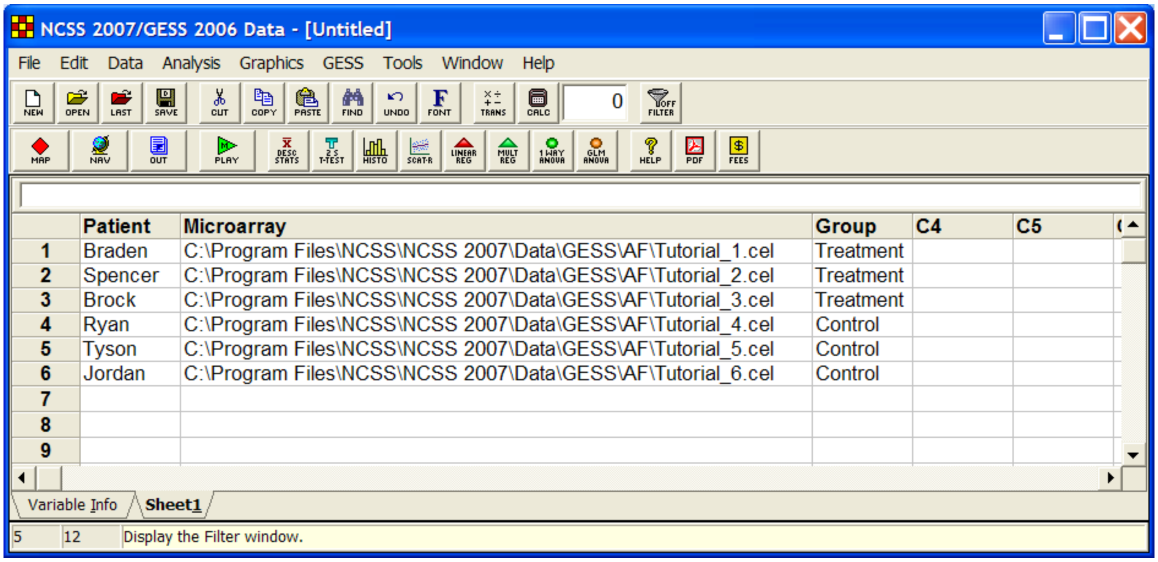
Tutorial

This chapter will quickly familiarize you with the most basic concepts and analysis steps required for a complete microarray analysis using ***GESS***. With the amount of data obtained in microarray experiments, the process of statistical analysis can be very frustrating. ***GESS*** has been designed to take the frustration out of microarray analysis with easy-to-use data entry, importing, pre-processing, and analysis routines. Following a general outline of the analysis steps, we will take you through a step-by-step analysis example, from beginning to end.

Basic Analysis Steps

The following is an outline of the steps involved in a microarray data analysis using ***GESS***. For detailed information about each step, see the tutorial below.

Step 1 – Enter the Microarray Data File Names and Experiment Information into the Spreadsheet


NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 OFF FILTER
MAP NAV OUT PLAY DESC STATS 2 S T-TEST HISTO SCAT-R LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info
Sheet1

	Patient	Microarray	Group	C4	C5
1	Braden	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel	Treatment		
2	Spencer	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel	Treatment		
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment		
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control		
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control		
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control		
7					
8					
9					

5 12 Display the Filter window.

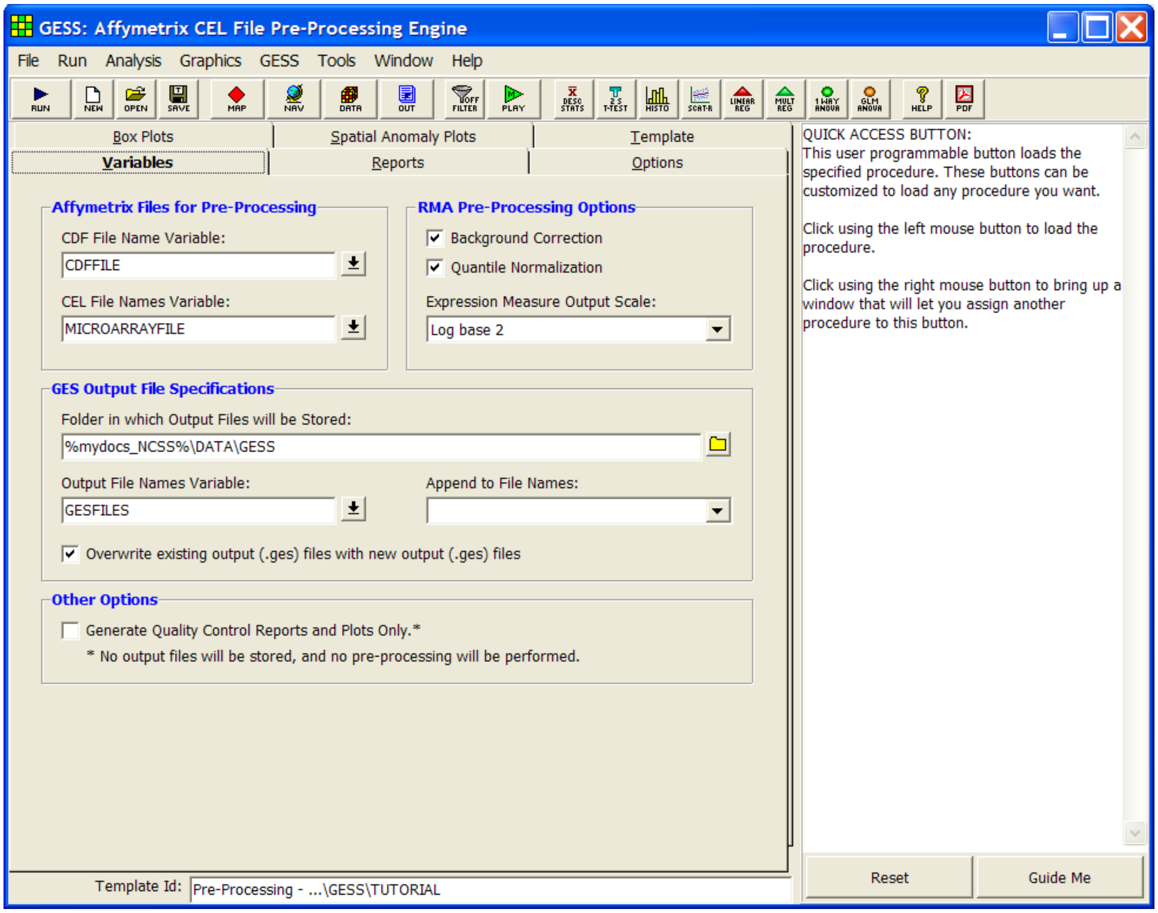
The data system in ***GESS*** is based on a familiar spreadsheet user-interface. Each individual in a microarray experiment is represented by a single row on the database. The microarray data file corresponding to each individual is listed by name in a single column. Data for other variables can be specified as necessary.

The file types that can be imported directly into ***GESS*** are

1. Affymetrix Intensity Files (.cel)
2. Affymetrix Expression Files (.chp)
3. Agilent Intensity Files (.txt)

4. Genepix Intensity Files (.gpr)
5. Generic Two-Channel Intensity Files (.txt)
6. Generic Expression Data Files (.txt, .csv, .dat, .dcp, etc.)

Step 2 – Import the Data Files into GESS and Create .ges Files to Be Used in Further Analyses



Affymetrix Files for Pre-Processing

CDF File Name Variable: CDFFILE

CEL File Names Variable: MICROARRAYFILE

RMA Pre-Processing Options

☑ Background Correction

☑ Quantile Normalization

Expression Measure Output Scale: Log base 2

GES Output File Specifications

Folder in which Output Files will be Stored: %mydocs_NCSS%\DATA\GESS

Output File Names Variable: GESFILES

Append to File Names:

☑ Overwrite existing output (.ges) files with new output (.ges) files

Other Options

□ Generate Quality Control Reports and Plots Only.*

* No output files will be stored, and no pre-processing will be performed.

QUICK ACCESS BUTTON:
This user programmable button loads the specified procedure. These buttons can be customized to load any procedure you want.

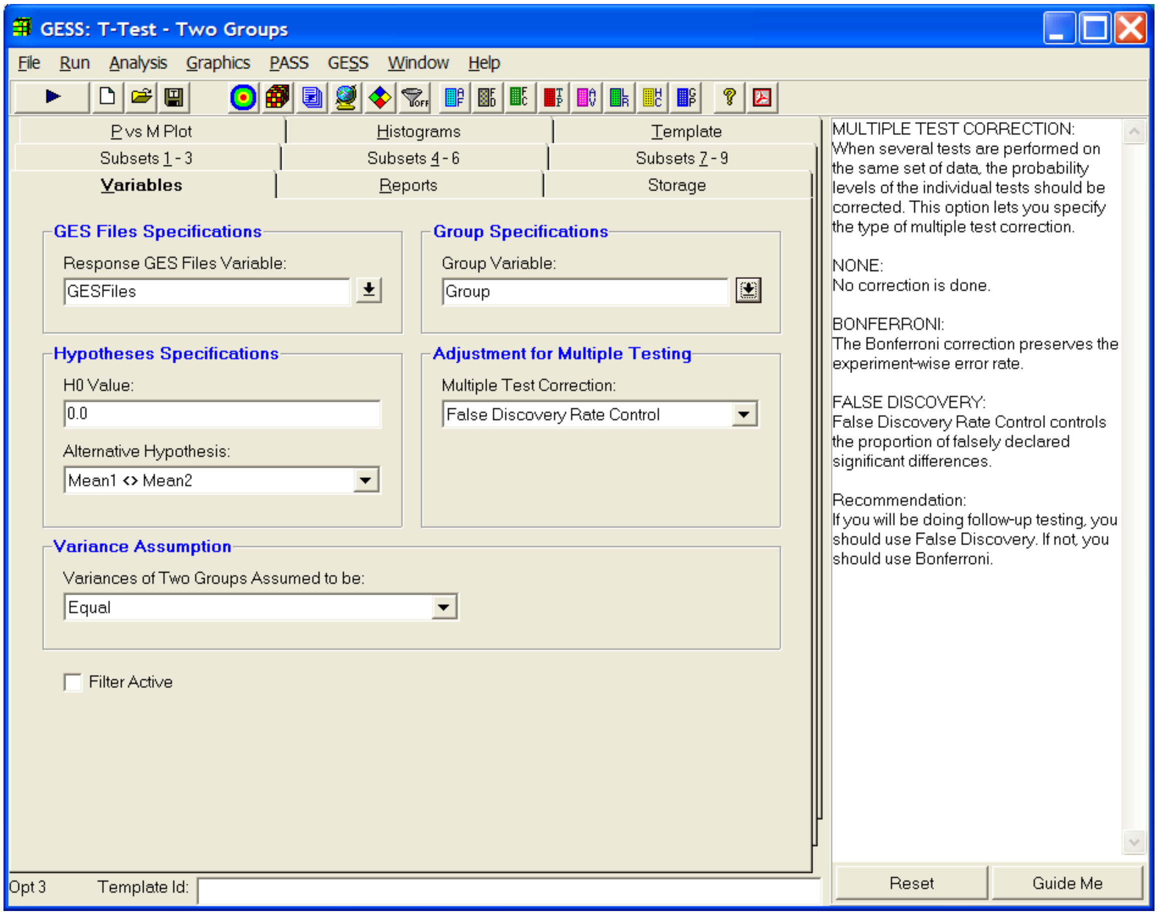
Click using the left mouse button to load the procedure.

Click using the right mouse button to bring up a window that will let you assign another procedure to this button.

Template Id: Pre-Processing - ...\GESS\TUTORIAL

Once the data and file names have been entered into the spreadsheet, you must convert the microarray files (.cel, .chp, .gpr, etc.) into .ges files, the file type used by ***GESS***. The .ges files are read much faster than the original files in subsequent analyses. In the case of Affymetrix .cel files, Agilent .txt files, and GenePix .gpr files, the importing process involves pre-processing as well. A separate import or pre-processing engine is available for each different import-file type.

Step 3 – Perform Statistical Analysis and Output Results



GES Files Specifications

Response GES Files Variable: GESFiles

Group Specifications

Group Variable: Group

Hypotheses Specifications

H0 Value: 0.0

Alternative Hypothesis: Mean1 <> Mean2

Adjustment for Multiple Testing

Multiple Test Correction: False Discovery Rate Control

Variance Assumption

Variances of Two Groups Assumed to be: Equal

□ Filter Active

Opt 3 Template Id:

MULTIPLE TEST CORRECTION:
When several tests are performed on the same set of data, the probability levels of the individual tests should be corrected. This option lets you specify the type of multiple test correction.

NONE:
No correction is done.

BONFERRONI:
The Bonferroni correction preserves the experiment-wise error rate.

FALSE DISCOVERY:
False Discovery Rate Control controls the proportion of falsely declared significant differences.

Recommendation:
If you will be doing follow-up testing, you should use False Discovery. If not, you should use Bonferroni.

After importing and pre-processing, you can proceed to the completion of statistical analysis. The statistical routines in ***GESS*** automatically adjust the p-values for multiple testing (if desired). To complete the analysis, simply fill out the procedure window and click Run. The output listing the significant genes and displaying powerful graphics will automatically be displayed. You can then save the data from significant genes to the spreadsheet for further analysis using ***NCSS***, or cut and paste the report directly into a document or presentation. You also have the option to specify subsets of genes for analysis, thus, allowing you to narrow the list of target genes and increase your power for detecting significant differences or effects.

Several statistical routines are available in ***GESS***, each with full documentation:

1. Fold-Change Analysis
2. Paired, One-Sample and Two-Sample T-Tests
3. Analysis of Variance (GLM)
4. Repeated Measures ANOVA
5. Cox Regression
6. Logistic Regression
7. Multiple Regression
8. Principal Components Analysis
9. Hierarchical Cluster Analysis

Tutorial

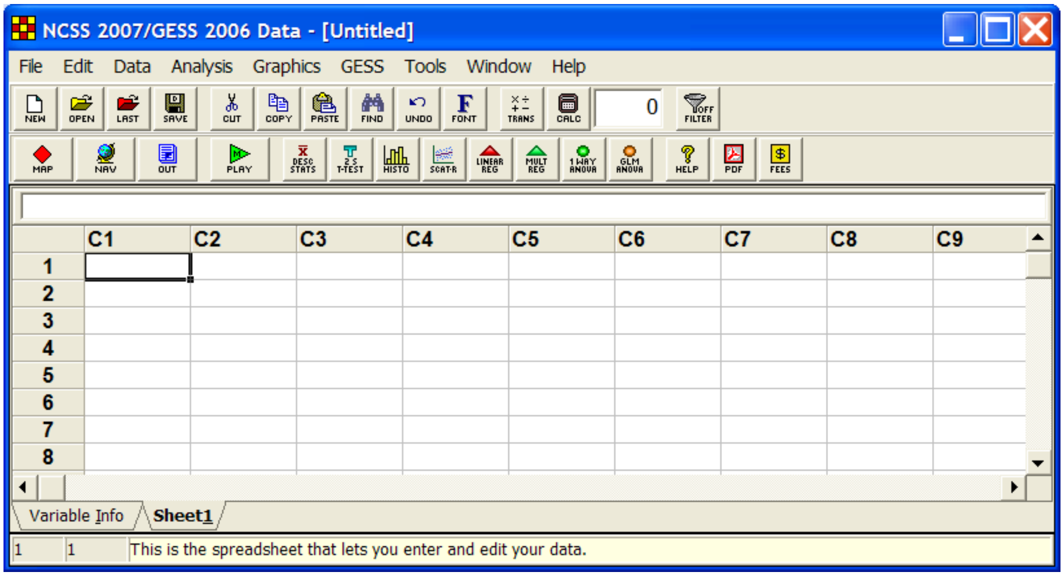
We will now take you through the analysis process step by step. In this analysis we will use a sample Affymetrix dataset consisting of six .cel files. Our goal will be to perform a T-test for differential expression and a follow-up cluster analysis on significant genes. This tutorial will show you how a general statistical analysis is done, from a blank spreadsheet to a final report. While the individual steps may vary for other microarray platforms, e.g. GenePix, Agilent, etc., the overall analysis process is the same. Specific information relating to each importing and analysis procedure can be found in other chapters of this manual.

Step 1 – Enter the Microarray Data File Names and Experiment Information into the Spreadsheet

We will first present the basic steps necessary to start ***GESS***, change the column names, and enter data and file names. The input files for this tutorial are stored in the *C:\Program Files\NCSS \NCSS 2007\Data\GESS\AF* directory. After completing this tutorial, your dataset should match the TUTORIAL dataset found in the *C:\...\[My] Documents\NCSS\NCSS 2007\Data\GESS* directory.

1 Launch GESS.

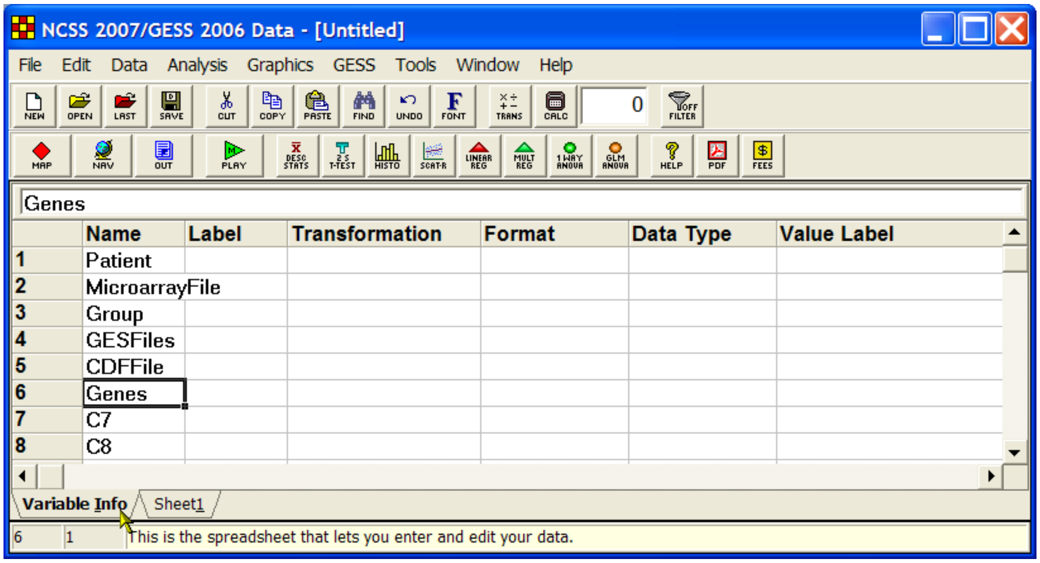
- Double click the ***GESS*** desktop icon or launch ***GESS*** from the Windows Start menu by clicking on **All Programs**. The GESS Data window will appear.



This is the spreadsheet that lets you enter and edit your data.

2 Change the Column Names.

- Click on the **Variable Info** tab in the bottom left-hand corner of the screen.
- In the **Name** column, select the cell containing **C1**, and enter **Patient**. Enter **MicroarrayFile**, **Group**, **GESFiles**, **CDFFile**, and **Genes** for **C2-C6**, respectively. We will fill in these variables as we go through this tutorial.
- Click on the **Sheet1 tab** at the bottom to return to the data sheet. The columns now have the new names as their titles.


NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 FILTER
MAP NAV OUT PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info
Sheet1

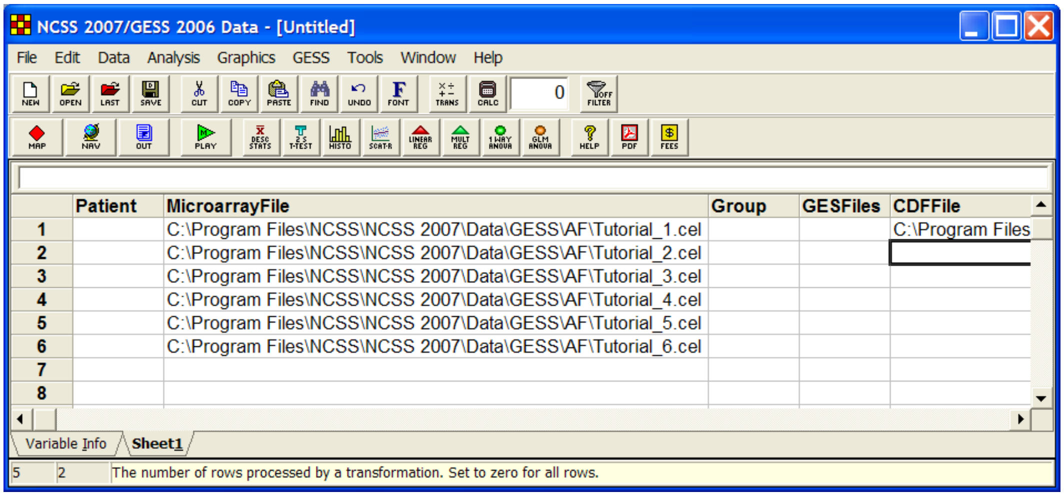
Genes

	Name	Label	Transformation	Format	Data Type	Value Label
1	Patient					
2	MicroarrayFile					
3	Group					
4	GESFiles					
5	CDFFile					
6	Genes					
7	C7					
8	C8					

6 1 This is the spreadsheet that lets you enter and edit your data.

3 Enter the Names of Existing Microarray Files for Analysis.

- Highlight the first cell in the **MicroarrayFile** column. Select **Edit**, then **Enter File Name**, or hit **F7** to launch the file browser. Under **Save as type** select **Affymetrix (*.CEL)**.
- Browse to the folder into which you installed ***GESS*** (usually *C:\Program Files\NCSS \NCSS 2007*).
- Select the **DATA** subdirectory of the **NCSS 2007** directory.
- Open the **GESS** folder and then the **AF** folder.
- Select the file **Tutorial_1.cel** and click **Save**. This will save the file name and path to the spreadsheet. Repeat the preceding steps for **Tutorial_2.cel through Tutorial_6.cel**. Hint: Using **F7** to launch the file browser makes this process go very fast.
- Highlight the first cell in the **CDFFile** column. Select **Edit**, then **Enter File Name**, or hit **F7** to launch the file browser. Under **Save as type** select **Affymetrix (*.CDF)**.
- Browse to the folder into which you installed ***GESS*** (usually *C:\Program Files\NCSS \NCSS 2007*).
- Select the **DATA** subdirectory of the **NCSS 2007** directory.
- Open the **GESS** folder and then the **AF** folder.
- Select the file **Test3.cdf** and click **Save**. This will save the file name and path to the spreadsheet.
  Note: The .cdf file is only required when importing Affymetrix .cel or .chp files, and must be obtained independently of the ***GESS*** system before you can perform your analysis. You can download Affymetrix .cdf library files from www.Affymetrix.com.
- Expand the Microarray_File and CDFFile column widths by choosing **Edit**, the Re**size Rows and Columns**, then **Resize using Data and Titles** from the Data window menus. This will allow you to see the entire file path when more data are entered.

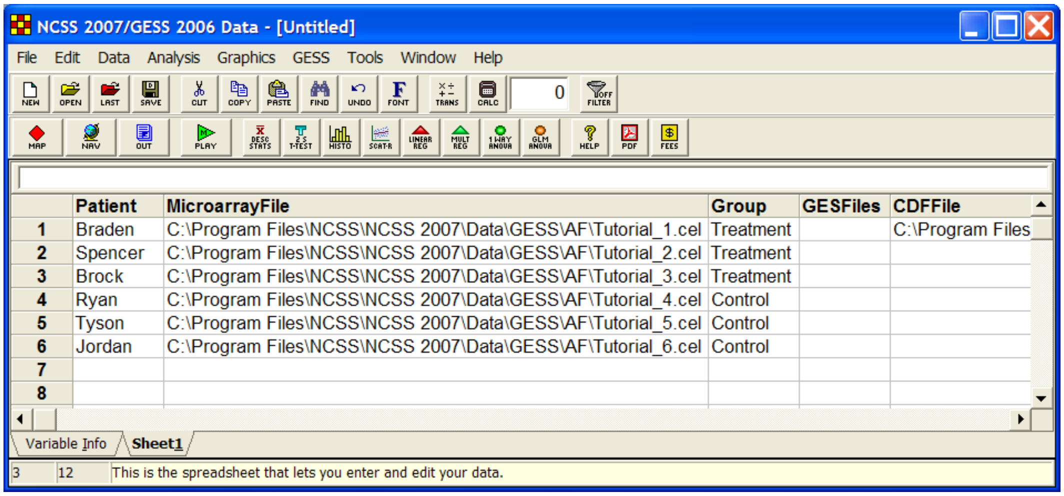

NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 FILTER
MAP NAV OUT PLAY DESC STATS T-TEST HISTO SCAT-R LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info
Sheet1

	Patient	MicroarrayFile	Group	GESFiles	CDFFile
1		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel			C:\Program Files
2		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel			
3		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel			
4		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel			
5		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel			
6		C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel			
7					
8					

5 2 The number of rows processed by a transformation. Set to zero for all rows.

4 Enter the Experiment Data.

- In the **Patient** column, enter **Braden**, **Spencer**, **Brock**, **Ryan**, **Tyson**, and **Jordan**, as the individual names.
- In the **Group** column, enter **Treatment** for the first three rows and **Control** for the last three rows.
- Save the data file as **TUTORIAL.S0** by clicking on **File** and then **Save**. The data are now ready for pre-processing and analysis.


NCSS 2007/GESS 2006 Data - [Untitled]
File Edit Data Analysis Graphics GESS Tools Window Help
NEW OPEN LAST SAVE CUT COPY PASTE FIND UNDO FONT TRANS CALC 0 FILTER
MAP NAV OUT PLAY DESC STATS T-TEST HISTO SCAT-R LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF FEES
Variable Info
Sheet1

	Patient	MicroarrayFile	Group	GESFiles	CDFFile
1	Braden	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel	Treatment		C:\Program Files
2	Spencer	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel	Treatment		
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment		
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control		
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control		
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control		
7					
8					

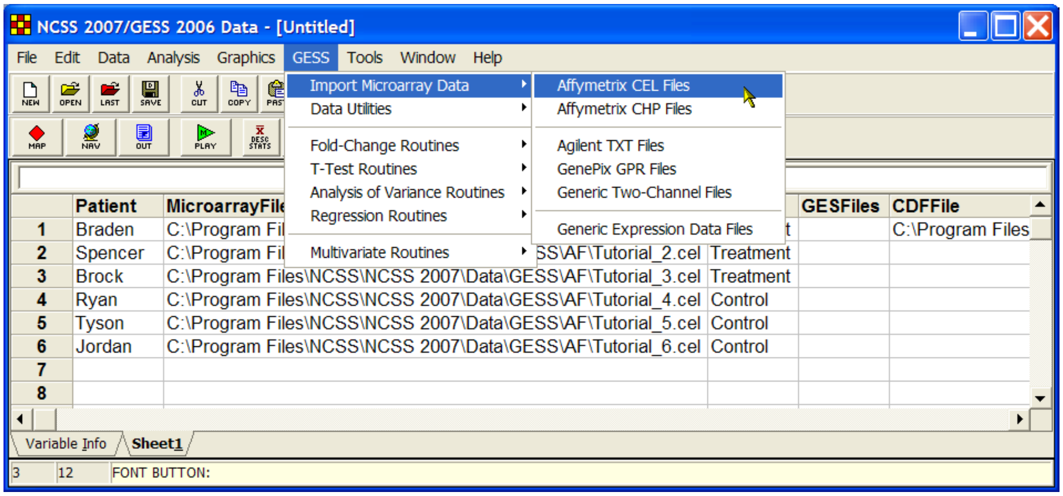
3 12 This is the spreadsheet that lets you enter and edit your data.

Step 2 – Import the Data Files into GESS and Create .ges Files to Be Used in Further Analyses

We now present the steps necessary to import files into ***GESS***. The input files for this tutorial are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF* directory. To run this example, take the following steps or load the **GESS Tutorial - Step 2** template on the Affymetrix CEL File Pre-Processing Engine Template tab.

1 Launch the Importing or Pre-Processing Procedure.

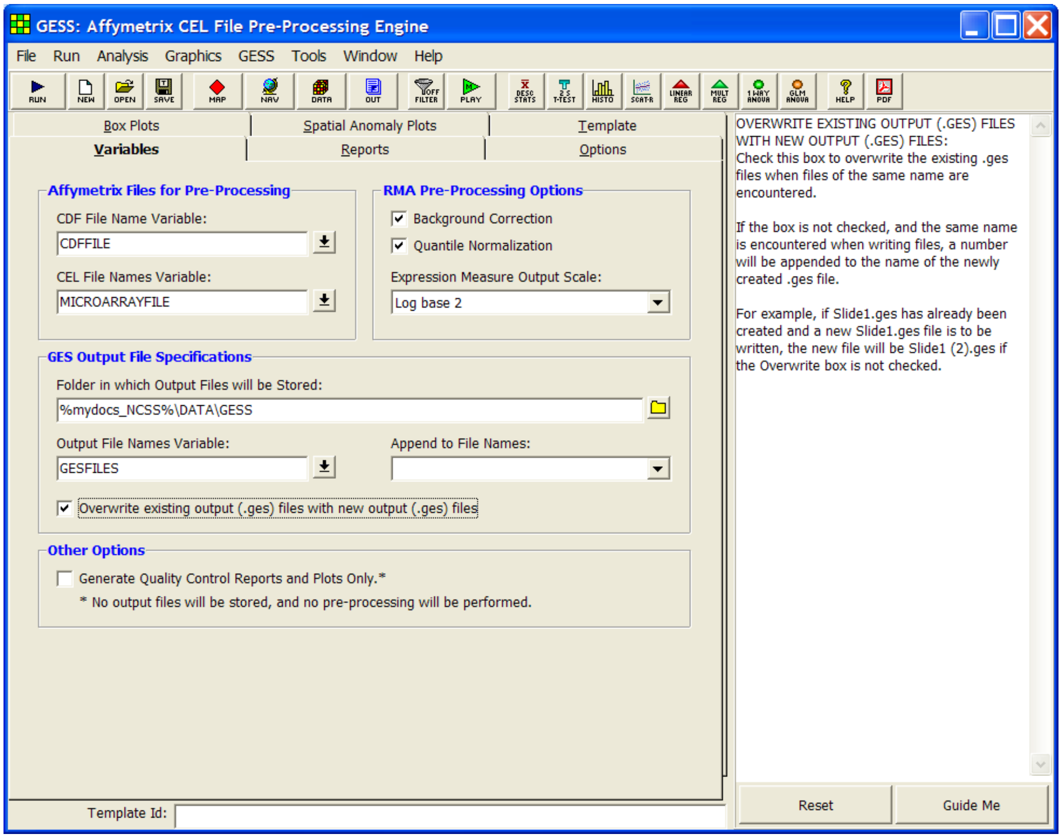
- On the menus, select **GESS**, then **Import Microarray Data**, then **Affymetrix CEL Files**. The Affymetrix CEL File Pre-Processing Engine procedure will be displayed. Hint: Right click on one of the eight quick-launch icons on the spreadsheet toolbar to select icons that you use most frequently. The new icons will be placed on the toolbar.
- On the Affymetrix CEL File Pre-Processing Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.



	Patient	MicroarrayFile		GESFiles	CDFFile
1	Braden	C:\Program Fil	t		C:\Program Files
2	Spencer	C:\Program Fil SS\AF\Tutorial_2.cel	Treatment		
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment		
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control		
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control		
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control		
7					
8					

2 Specify the Variables.

- On the Affymetrix CEL File Pre-Processing Engine window, select the **Variables tab**.
- Set the **CDF File Name Variable** to **CDFFile**.
- Set the **CEL File Names Variable** to **MicroarrayFile**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**. The designator "%mydocs_NCSS%" represents the path to the ***NCSS*** personal data folder (commonly *C:\...\[My] Documents\NCSS\NCSS 2007*).
- Set the **Output File Names Variable** to **GESFiles**.
- Check the box next to **Overwrite existing output (.ges) files with new output (.ges) files**.
- Leave all other options under the Variables tab at their default settings.



Affymetrix Files for Pre-Processing

CDF File Name Variable: CDFFILE

CEL File Names Variable: MICROARRAYFILE

RMA Pre-Processing Options

☑ Background Correction

☑ Quantile Normalization

Expression Measure Output Scale: Log base 2

GES Output File Specifications

Folder in which Output Files will be Stored: %mydocs_NCSS%\DATA\GESS

Output File Names Variable: GESFILES

Append to File Names:

☑ Overwrite existing output (.ges) files with new output (.ges) files

Other Options

☐ Generate Quality Control Reports and Plots Only.*

* No output files will be stored, and no pre-processing will be performed.

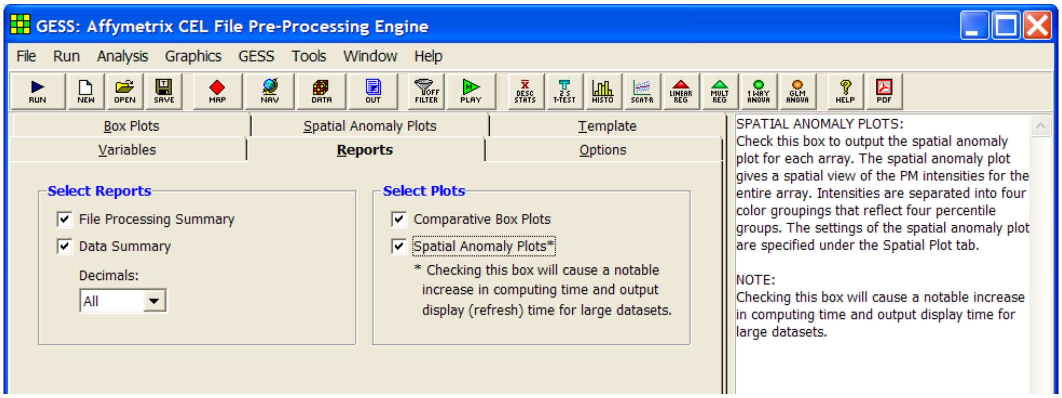
OVERWRITE EXISTING OUTPUT (.GES) FILES WITH NEW OUTPUT (.GES) FILES:
Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file.

For example, if Slide1.ges has already been created and a new Slide1.ges file is to be written, the new file will be Slide1 (2).ges if the Overwrite box is not checked.

3 Specify the Reports.

- Select the **Reports tab**.
- Check the box next to **Spatial Anomaly Plots**.
- Leave all other options under the Reports tab and other tabs at their default settings.



Select Reports

☑ File Processing Summary

☑ Data Summary

Decimals: All

Select Plots

☑ Comparative Box Plots

☑ Spatial Anomaly Plots*

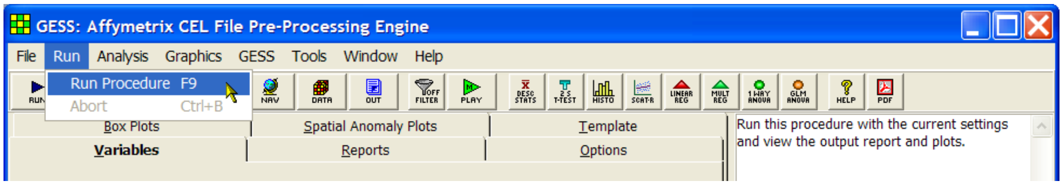
* Checking this box will cause a notable increase in computing time and output display (refresh) time for large datasets.

SPATIAL ANOMALY PLOTS:
Check this box to output the spatial anomaly plot for each array. The spatial anomaly plot gives a spatial view of the PM intensities for the entire array. Intensities are separated into four color groupings that reflect four percentile groups. The settings of the spatial anomaly plot are specified under the Spatial Plot tab.

NOTE:
Checking this box will cause a notable increase in computing time and output display time for large datasets.

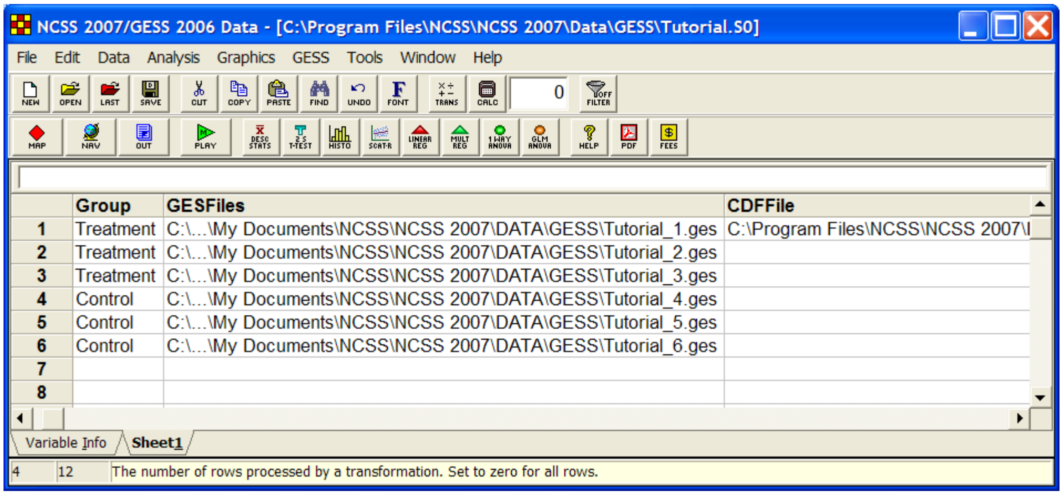
4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the toolbar at the top).



Run this procedure with the current settings and view the output report and plots.

- The **GESFiles** column on the spreadsheet now appears as follows (after resizing the column width):



	Group	GESFiles	CDFFile
1	Treatment	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_1.ges	C:\Program Files\NCSS\NCSS 2007\I
2	Treatment	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_2.ges	
3	Treatment	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_3.ges	
4	Control	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_4.ges	
5	Control	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_5.ges	
6	Control	C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_6.ges	
7			
8			

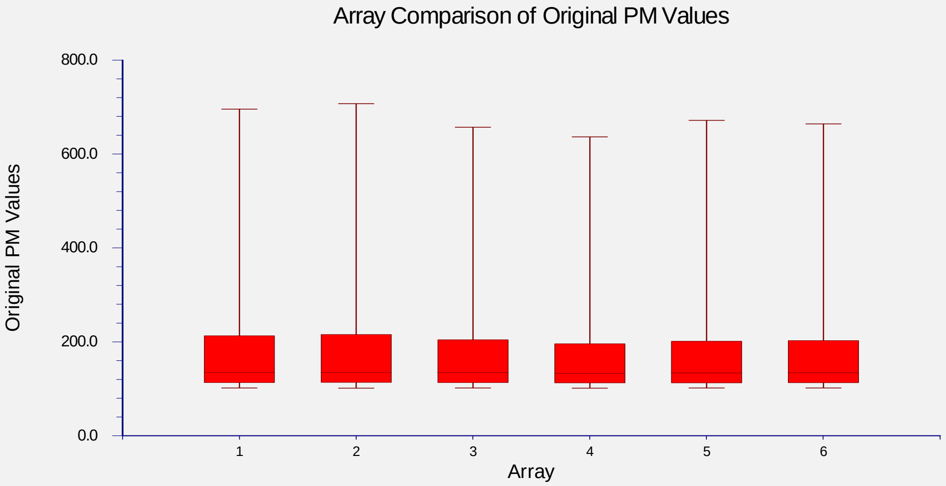
4 12 The number of rows processed by a transformation. Set to zero for all rows.

Notice that the **GESFiles** column has been automatically filled with the names of the newly created .ges files. These .ges files will be used by ***GESS*** when array data is needed in future analyses. The report output is given below.

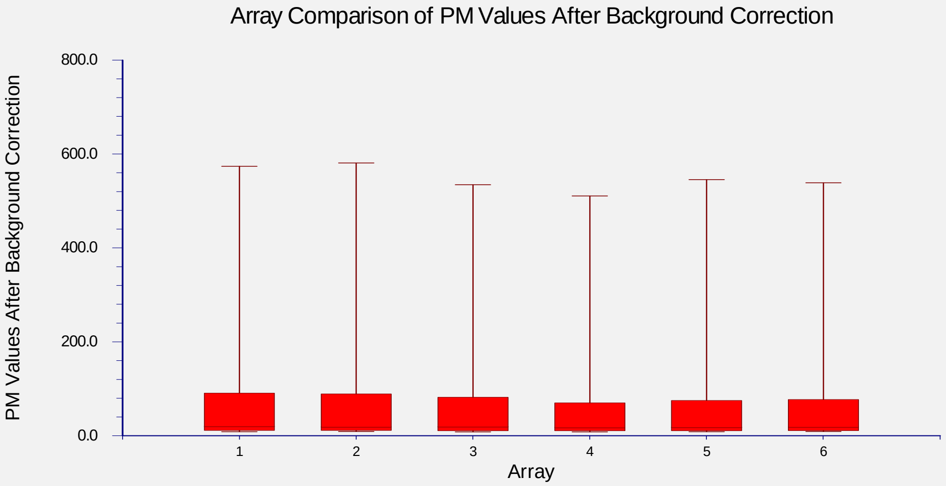
Report Output

```
Input File Summary

                 Number of
Row              Probes           File Name
1                15876            C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_1.cel
2                15876            C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_2.cel
3                15876            C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel
4                15876            C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel
5                15876            C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel
6                15876            C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel

CDF File Name: C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Test3.cdf
CDH File Name: C:\...\My Documents\NCSS\NCSS 2007\Data\GESS\CDH\Test3.cdh


Pre-Processing Methods Summary

Task Name                    Method Selected
Background Correction        RMA (Model-Based)
Normalization                Quantile
Summarization                Median Polish
Output Scale                 Log base 2


Output File Summary

                 Number of
Row              Probe Sets       File Name
1                345              C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_1.ges
2                345              C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_2.ges
3                345              C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_3.ges
4                345              C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_4.ges
5                345              C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_5.ges
6                345              C:\...\My Documents\NCSS\NCSS 2007\DATA\GESS\Tutorial_6.ges


Numerical Summary of Original PM Intensities

                             10th          25th          50th          75th          90th
Row      Minimum     Percentile    Percentile    Percentile    Percentile    Percentile       Maximum
1           61.5          101.8         113.5           136         212.8        695.65         27499
2           64.3          101.3         113.8         135.8         215.5           707       27151.3
3           70.5         101.65         113.3         135.5         204.3         656.8       23705.8
4           63.3          101.5         112.8         133.8         195.8        636.55       37486.5
5           65.3          101.8         112.8         134.8         201.3        671.65       31738.5
6             66         101.65           113         135.3         202.8         664.3         44675


Numerical Summary of Expression Values (Log2 Scale)

                             10th          25th          50th          75th          90th
Row      Minimum     Percentile    Percentile    Percentile    Percentile    Percentile       Maximum
1       3.432443       3.777095      4.028214       4.30166      4.749565      5.163875       9.09646
2        3.37914       3.819739      4.030816      4.307514      4.677792      5.144782      9.067124
3       3.472117       3.834185      4.031287      4.257486      4.641731      5.165857      9.163662
4       3.358151       3.846722      4.037156      4.325556      4.664896      5.110379      9.000655
5       3.435157       3.892049      4.071929      4.310611      4.635222      5.114842      9.055325
6       3.243854       3.893716       4.08064      4.341212      4.611947      5.037131      8.962646
```

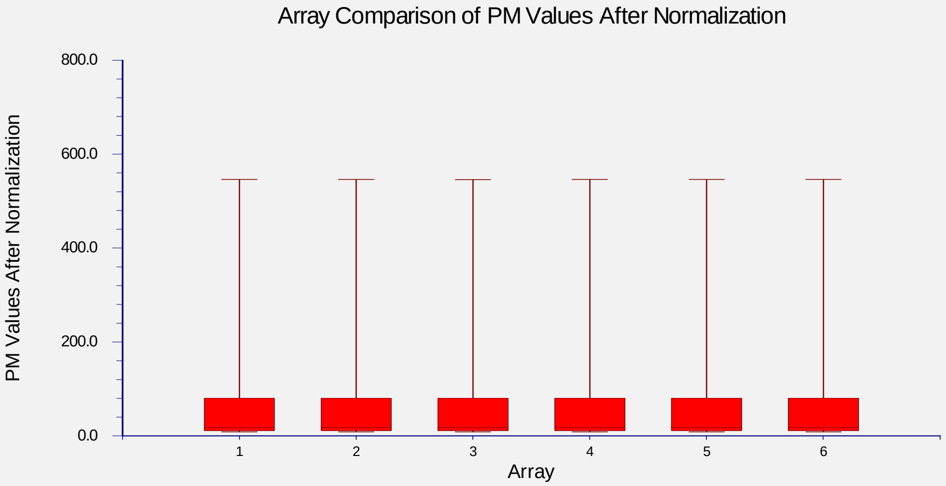
Box Plot Section



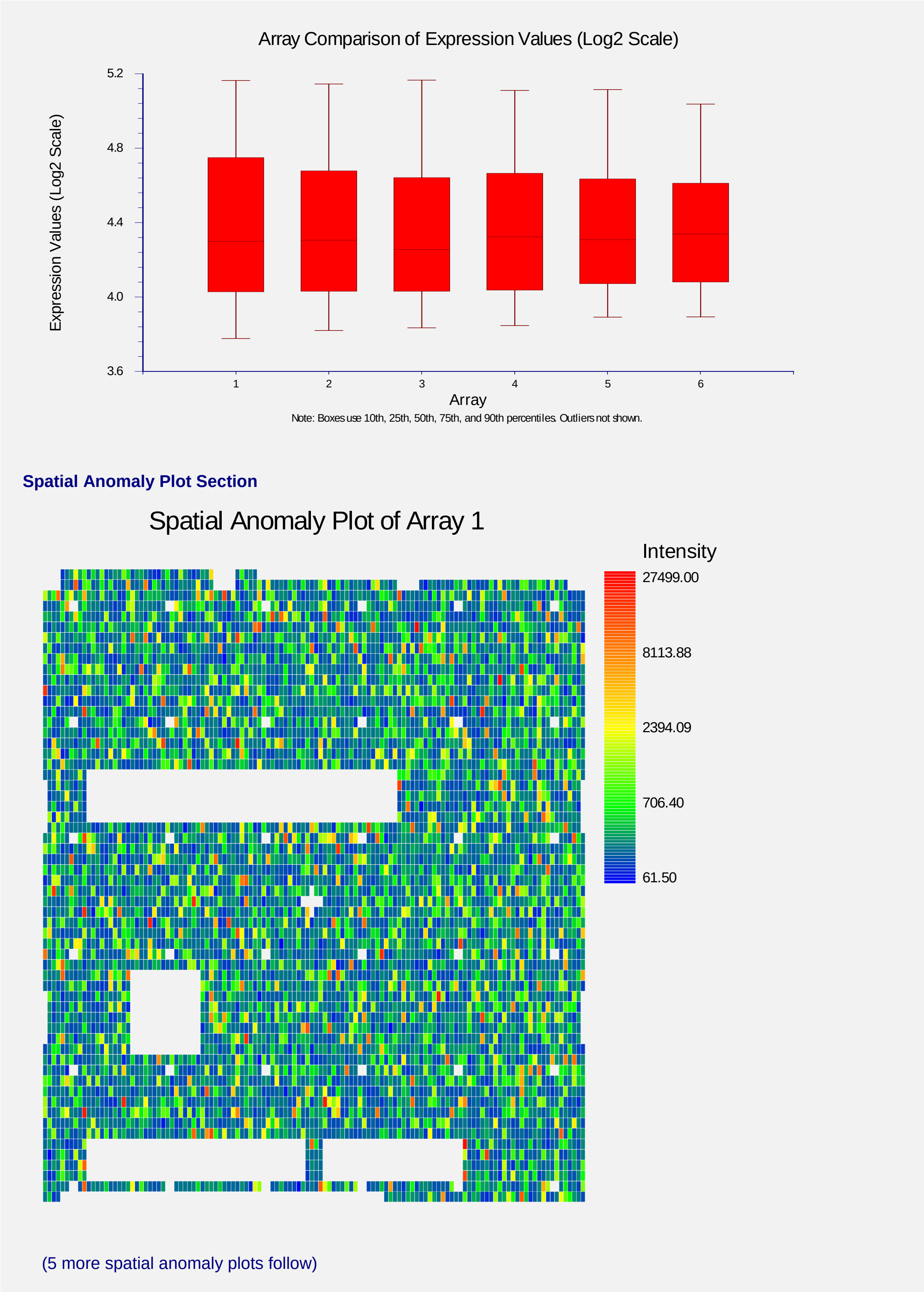
Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.


Array Comparison of Expression Values (Log2 Scale)
Expression Values (Log2 Scale)
5.2
4.8
4.4
4.0
3.6
1
2
3
4
5
6
Array
Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Spatial Anomaly Plot Section

Spatial Anomaly Plot of Array 1
Intensity
27499.00
8113.88
2394.09
706.40
61.50

(5 more spatial anomaly plots follow)

The reports give you a summary of the files processed and the data they contain, as well as graphical representations of the individual array data. Double-click on any plot to enlarge it.

Step 3 – Perform Statistical Analyses and Output Results

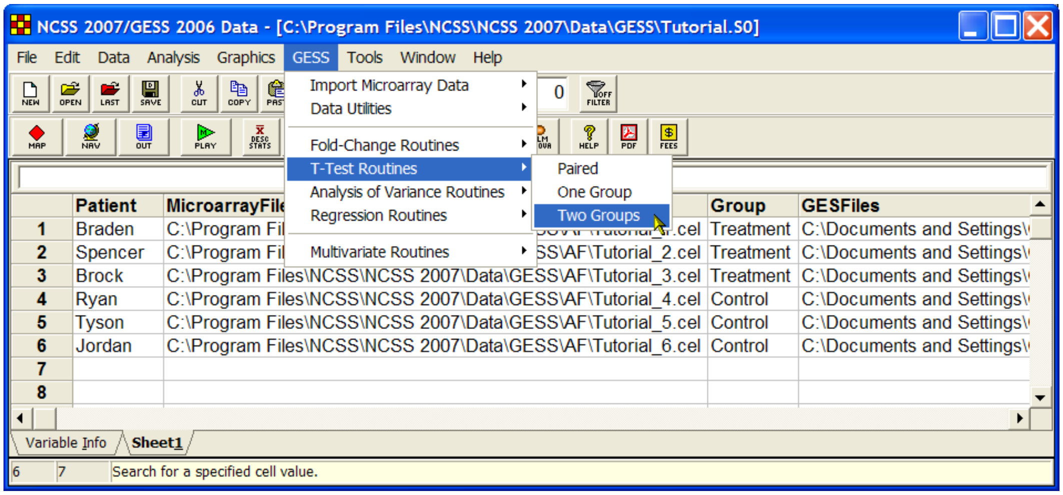
We now present some basic statistical analyses in ***GESS***. We will analyze the data using a two-sample T-test, followed by a hierarchical cluster analysis of significant genes.

Two-Sample T-Test Steps

To run this example, take the following steps or load the **GESS Tutorial - Step 3** template on the T-Test - Two Groups Template tab.

1 **Launch the T-Test - Two Groups Procedure.**

- On the menus, select **GESS**, then **T-Test Routines**, then **Two Groups**. The T-Test - Two Groups procedure will be displayed. Hint: Right click on one of the eight quick-launch icons on the spreadsheet toolbar to select icons that you use most frequently. The new icons will be placed on the toolbar.
- On the T-Test - Two Groups window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

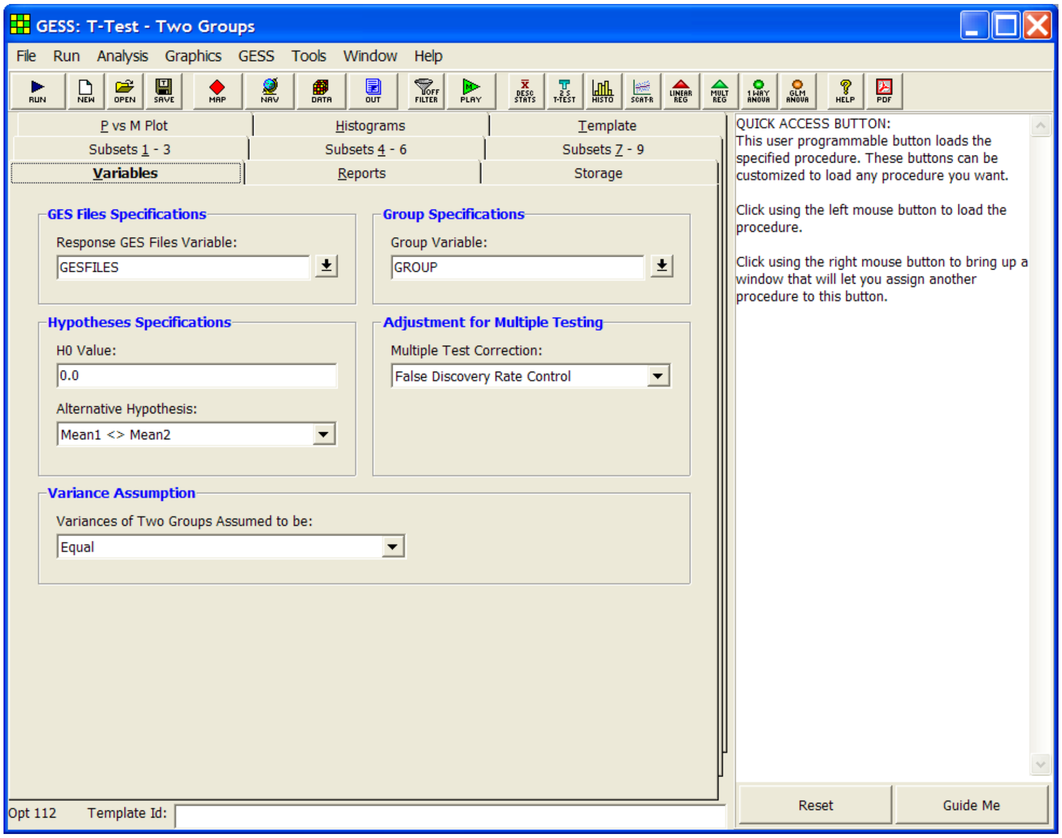


	Patient	MicroarrayFile	Group	GESFiles
1	Braden	C:\Program Fil … .cel	Treatment	C:\Documents and Settings\
2	Spencer	C:\Program Fil … SS\AF\Tutorial_2.cel	Treatment	C:\Documents and Settings\
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_3.cel	Treatment	C:\Documents and Settings\
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control	C:\Documents and Settings\
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control	C:\Documents and Settings\
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control	C:\Documents and Settings\
7				
8				

6 7 Search for a specified cell value.

2 **Specify the Variables.**

- On the T-Test - Two Groups window, select the **Variables tab**.
- Set the **Response GES Files Variable** to **GESFiles**.
- Set the **Group Variable** to **Group**.
- Leave all other options under the Variables tab at their default settings.


GESS: T-Test - Two Groups
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
P vs M Plot
Histograms
Template
Subsets 1 - 3
Subsets 4 - 6
Subsets 7 - 9
Variables
Reports
Storage
Reset
Guide Me
Opt 112
Template Id:

GES Files Specifications

Response GES Files Variable: GESFILES

Group Specifications

Group Variable: GROUP

Hypotheses Specifications

H0 Value: 0.0

Alternative Hypothesis: Mean1 <> Mean2

Adjustment for Multiple Testing

Multiple Test Correction: False Discovery Rate Control

Variance Assumption

Variances of Two Groups Assumed to be: Equal

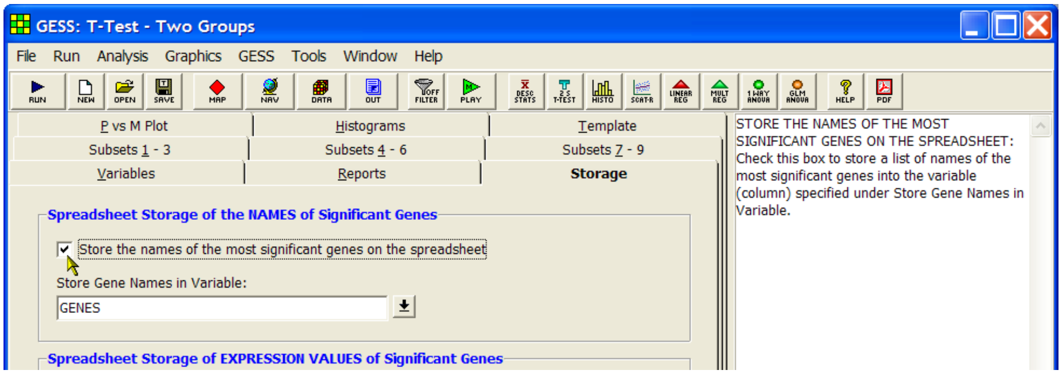
QUICK ACCESS BUTTON:
This user programmable button loads the specified procedure. These buttons can be customized to load any procedure you want.

Click using the left mouse button to load the procedure.

Click using the right mouse button to bring up a window that will let you assign another procedure to this button.

3 Specify the Storage Data.

- Select the **Storage tab**.
- Check the box next to **Store the names of the most significant genes on the spreadsheet**.
- Set the **Store Gene Names in Variable** to **Genes**.
- Leave all other options under the Storage tab and other tabs at their default settings.


GESS: T-Test - Two Groups
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
P vs M Plot
Histograms
Template
Subsets 1 - 3
Subsets 4 - 6
Subsets 7 - 9
Variables
Reports
Storage

Spreadsheet Storage of the NAMES of Significant Genes

☑ Store the names of the most significant genes on the spreadsheet

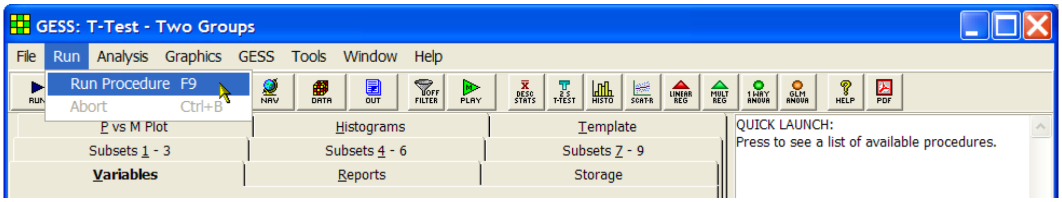
Store Gene Names in Variable: GENES

Spreadsheet Storage of EXPRESSION VALUES of Significant Genes

STORE THE NAMES OF THE MOST SIGNIFICANT GENES ON THE SPREADSHEET:
Check this box to store a list of names of the most significant genes into the variable (column) specified under Store Gene Names in Variable.

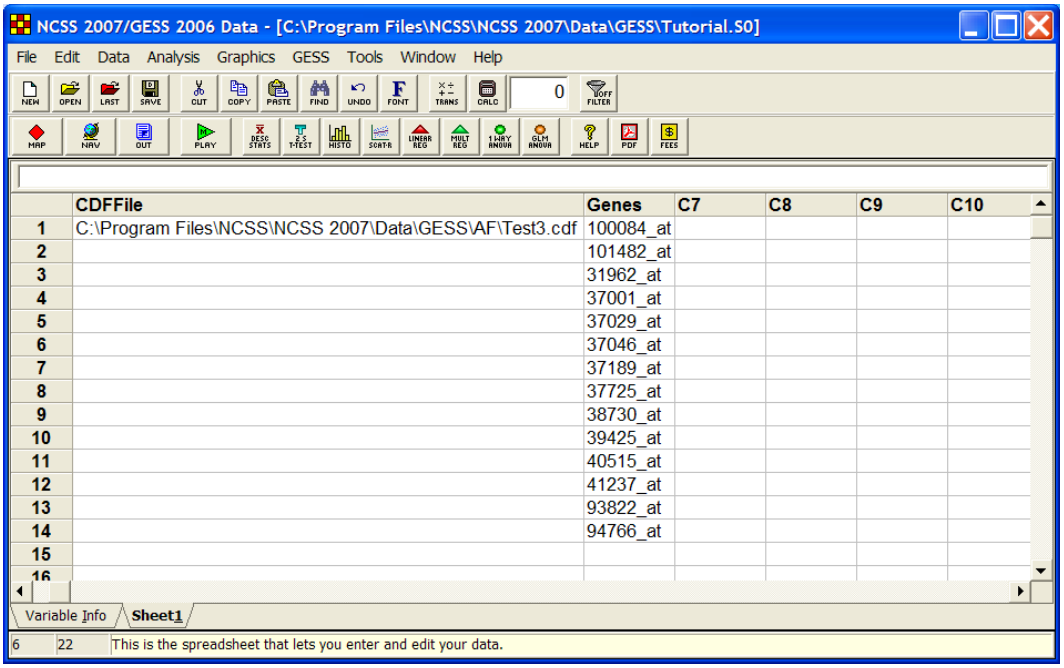
4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the toolbar at the top).



QUICK LAUNCH:
Press to see a list of available procedures.

The **Genes** column on the spreadsheet now appears as follows (after resizing the column width):



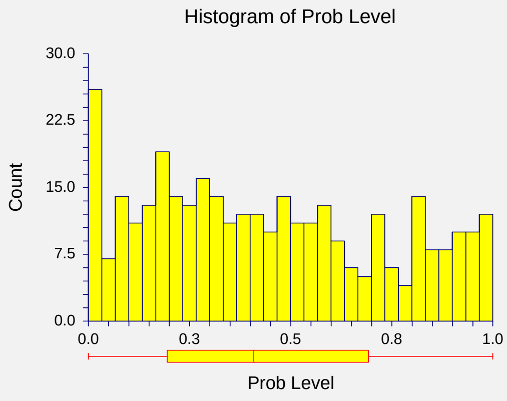
	CDFFile	Genes	C7	C8	C9	C10
1	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Test3.cdf	100084_at				
2		101482_at				
3		31962_at				
4		37001_at				
5		37029_at				
6		37046_at				
7		37189_at				
8		37725_at				
9		38730_at				
10		39425_at				
11		40515_at				
12		41237_at				
13		93822_at				
14		94766_at				
15						
16						

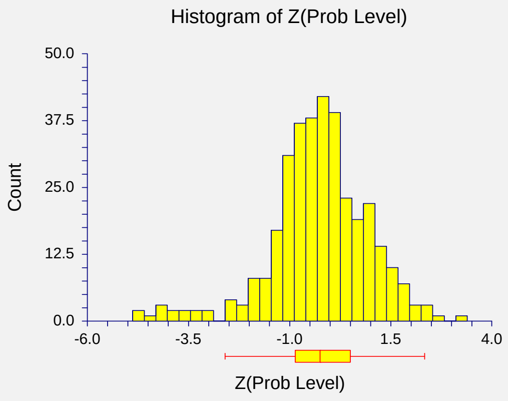
6 22 This is the spreadsheet that lets you enter and edit your data.

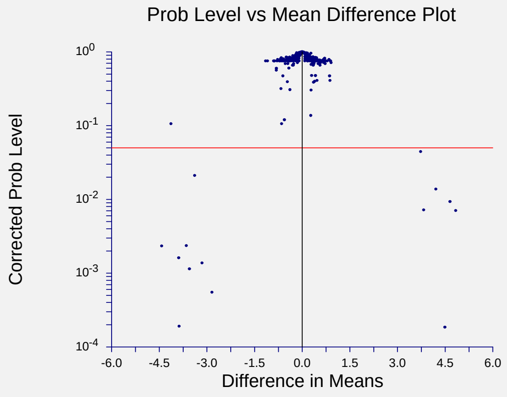
Notice that the **Genes** column has been automatically filled with the names of the most significant genes. Information about the T-test is given in the report. These genes will be used later for hierarchical clustering.

T-Test Output

```
T-Test Detail in Probability Level Order
Alternative Hypothesis: Mean of Control - Mean of Treatment <> 0

                                 FDR
                            Adjusted
                            Multiple          Single
Gene        Subset             Tests            Test                Counts        Mean   Standard
Name        Name          Prob Level      Prob Level   T Value     (N1/N2)  Difference      Error
41237_at    Other          0.0001855       0.0000005    57.770         3/3      4.4944     0.0778
101482_at   Other          0.0001912       0.0000011   -48.203         3/3     -3.8749     0.0804
94766_at    Other          0.0005509       0.0000048   -33.403         3/3     -2.8402     0.0850
37001_at    Other          0.0011430       0.0000133   -25.876         3/3     -3.5477     0.1371
37046_at    Other          0.0013776       0.0000200   -23.342         3/3     -3.1501     0.1350
31962_at    Other          0.0016171       0.0000281   -21.414         3/3     -3.8857     0.1815
37029_at    Other          0.0023416       0.0000475   -18.763         3/3     -4.4225     0.2357
38730_at    Other          0.0023667       0.0000549   -18.092         3/3     -3.6455     0.2015
100084_at   Other          0.0070727       0.0001845    13.304         3/3      4.8340     0.3633
93822_at    Other          0.0072033       0.0002088    12.892         3/3      3.8257     0.2968
40515_at    Other          0.0093642       0.0002986    11.766         3/3      4.6568     0.3958
37725_at    Other          0.0138263       0.0004809    10.410         3/3      4.2110     0.4045
39425_at    Other          0.0211671       0.0007976    -9.133         3/3     -3.3847     0.3706
37189_at    Other          0.0445618       0.0018083     7.368         3/3      3.7324     0.5066

Total number of hypothesis tests conducted = 345

Histograms and Plots Section
```







The basic report gives you detailed list of the most significant genes, along with histograms and a volcano plot showing the selected probability level cutoff of 0.05. Notice that the list of genes listed in the **Genes** column on the spreadsheet is the same as the list in this report. There are 14 genes that are differentially expressed, based on this analysis.

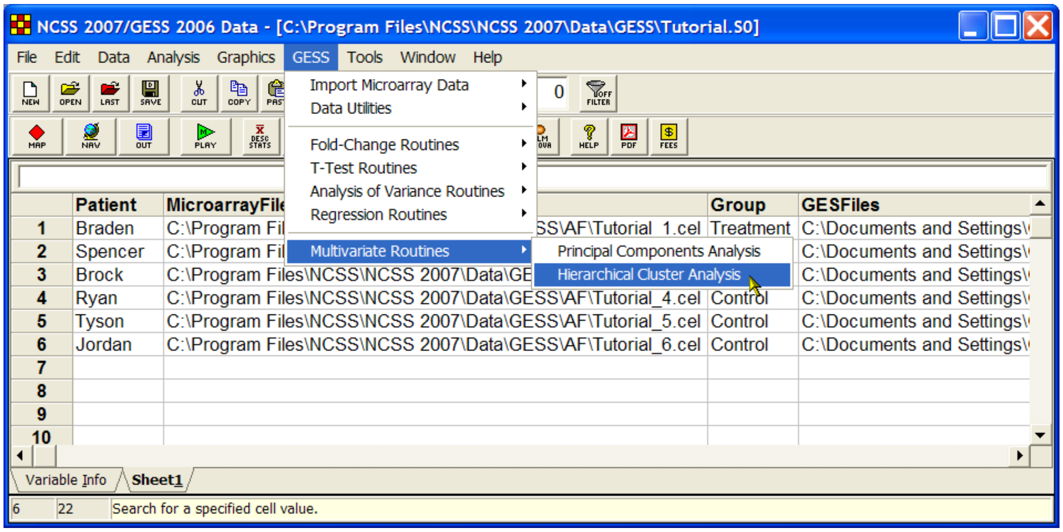
These reports and plots can be copied and pasted directly into a report or presentation. Double-click on any plot to enlarge it.

Hierarchical Cluster Analysis Steps

To run this example, take the following steps or load the **GESS Tutorial - Step 3** template on the Hierarchical Cluster Analysis Template tab.

1 Launch the Hierarchical Cluster Analysis Procedure.

- On the menus, select **GESS**, then **Multivariate Routines**, then **Hierarchical Cluster Analysis**. The Hierarchical Cluster Analysis procedure will be displayed. Hint: Right click on one of the eight quick-launch icons on the spreadsheet toolbar to select icons that you use most frequently. The new icons will be placed on the toolbar.
- On the Hierarchical Cluster Analysis window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

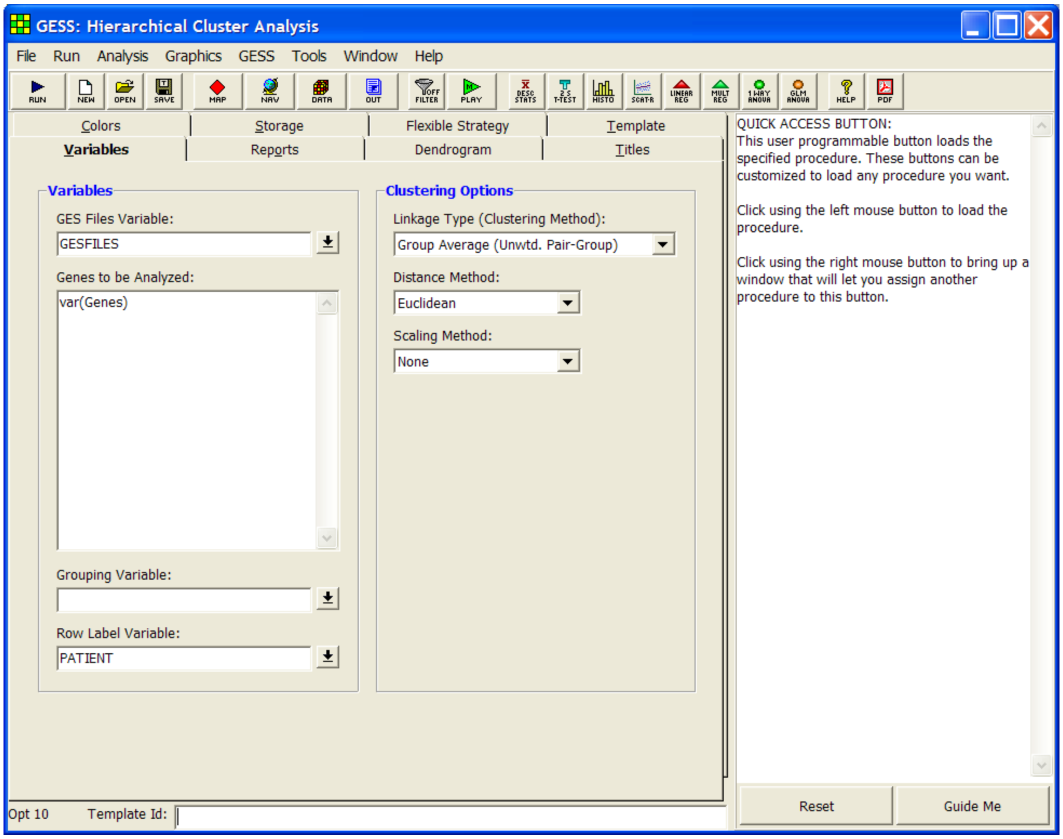


	Patient	MicroarrayFile	Group	GESFiles
1	Braden	C:\Program Fil … SS\AF\Tutorial_1.cel	Treatment	C:\Documents and Settings\
2	Spencer	C:\Program Fil		C:\Documents and Settings\
3	Brock	C:\Program Files\NCSS\NCSS 2007\Data\GE		C:\Documents and Settings\
4	Ryan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_4.cel	Control	C:\Documents and Settings\
5	Tyson	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_5.cel	Control	C:\Documents and Settings\
6	Jordan	C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF\Tutorial_6.cel	Control	C:\Documents and Settings\
7				
8				
9				
10				

6 22 Search for a specified cell value.

2 Specify the Variables.

- On the Hierarchical Cluster Analysis window, select the **Variables tab**.
- Set the **GES Files Variable** to **GESFiles**.
- Under **Genes to be Analyzed** enter **var(Genes)**.
- Set the **Row Label Variable** to **Patient**.
- Leave all other options under the Variables tab and other tabs at their default settings.


GESS: Hierarchical Cluster Analysis
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
Colors Storage Flexible Strategy Template
Variables Reports Dendrogram Titles
Variables
Clustering Options
Reset
Guide Me
Opt 10

GES Files Variable: GESFILES

Genes to be Analyzed: var(Genes)

Grouping Variable:

Row Label Variable: PATIENT

Linkage Type (Clustering Method): Group Average (Unwtd. Pair-Group)

Distance Method: Euclidean

Scaling Method: None

Template Id:

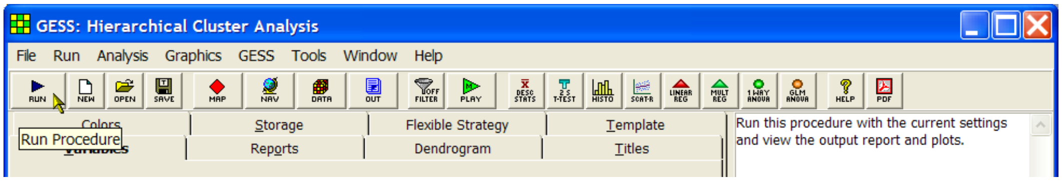
QUICK ACCESS BUTTON:
This user programmable button loads the specified procedure. These buttons can be customized to load any procedure you want.

Click using the left mouse button to load the procedure.

Click using the right mouse button to bring up a window that will let you assign another procedure to this button.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the toolbar at the top).

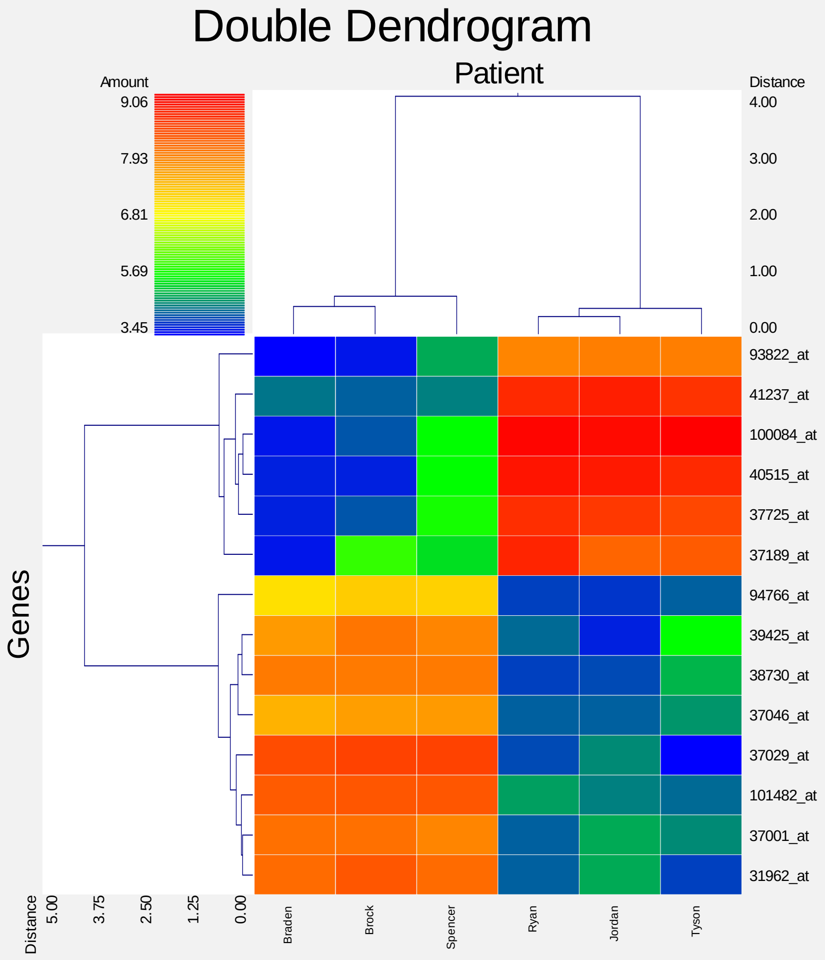

GESS: Hierarchical Cluster Analysis
File Run Analysis Graphics GESS Tools Window Help
RUN NEW OPEN SAVE MAP NAV DATA OUT FILTER PLAY DESC STATS T-TEST HISTO SCATR LINEAR REG MULT REG 1 WAY ANOVA GLM ANOVA HELP PDF
Colors Storage Flexible Strategy Template
Run Procedure
Variables Reports Dendrogram Titles

Run this procedure with the current settings and view the output report and plots.

Hierarchical Clustering Output

```
Clustering Method   Group Average (Unweighted Pair-Group)
Distance Type       Euclidean
Scale Type          None

Cluster Detail Section when Clustering Rows

Cluster     Rows in this Cluster
1           Braden, Brock
2           Ryan, Tyson, Jordan
None        Spencer

Cluster Detail Section when Clustering Genes

Cluster     Genes in Cluster
1           31962_at, 37001_at, 37029_at, 37046_at, 38730_at, 39425_at, 101482_at
2           37189_at, 37725_at, 40515_at, 41237_at, 100084_at, 93822_at
None        94766_at
```

Dendrogram Section



The basic report gives you a list of the clusters and a double dendrogram. The apparent separation in patient clusters corresponds exactly with the two treatment groups.

These reports and plots can be copied and pasted directly into a report or presentation. Double-click on the double dendrogram to enlarge it.

Chapter 110

Affymetrix CEL File Pre-Processing Engine

Introduction

The main purpose of this chapter is to describe the process of obtaining trusted relative expression values from Affymetrix GeneChip® output using the ***GESS*** Affymetrix CEL File Pre-Processing Engine. Following a brief background to the concept of microarrays, this chapter discusses many of the principles and practical aspects of high-density oligonucleotide arrays, including array production, image analysis, array and spot quality, background correction, normalization, and expression measure summarization. The chapter concludes with a tutorial of the entire process of using the Affymetrix CEL File Pre-Processing Engine to obtain expression values from GeneChip® output.

Chapter Structure

Overview

An overview of microarray concepts is presented first. This section is designed to familiarize a non-biologist with the concepts of gene expression and GeneChip® microarray design and construction.

Five Steps to Obtain Relative Expression Values

The background is followed by a summary of the five steps required to obtain final relative expression values for Affymetrix high-density oligonucleotide arrays, which can then be used in comparison analysis.

Step 1 - Hybridization - A sample is prepared, labeled, and introduced onto the prefabricated array. The sequences that are complementary to each probe will bind to the probe sequences.

Step 2 - Image Construction - The array is scanned producing an image file. The image file contains a value (representing a shade of gray) for every pixel on the array.

Step 3 - Image Processing - Image processing software is used to assign a single intensity score for each probe cell on the array.

Step 4 - Whole-Array Quality Assessment - Specialized plots and numeric summaries of the whole array indicate whether values obtained from the array can be trusted for comparison.

Step 5 - Expression Index Calculation (RMA) - The individual intensity values for each probe set are summarized into a single expression index. The summarization algorithm, RMA, involves three important steps:

1. Background Correction - A model-based algorithm is used to estimate the expression signal in the presence of background noise.
2. Normalization - Quantile normalization may be used to allow for accurate comparison between arrays. If normalization is not performed, interesting differences in expression may be obscured by variation arising from small differences in sample preparation, array production, and scanner processing.
3. Summarization - The corrected intensities are summarized into a single expression measure using a robust algorithm, median polish.

The result of these five steps is a set of trusted relative expression values that can be compared to corresponding values from other arrays in the experiment.

Affymetrix Files

The results file (.cel), obtained from Affymetrix image processing software and the chip description file (.cdf), necessary for interpretation of the .cel file are described.

Entering CDF and CEL Files

Details for entering .cel and .cdf files into the spreadsheet are given.

Procedure Options

The options available in ***GESS*** for pre-processing .cel files are described in detail.

Tutorials / Examples

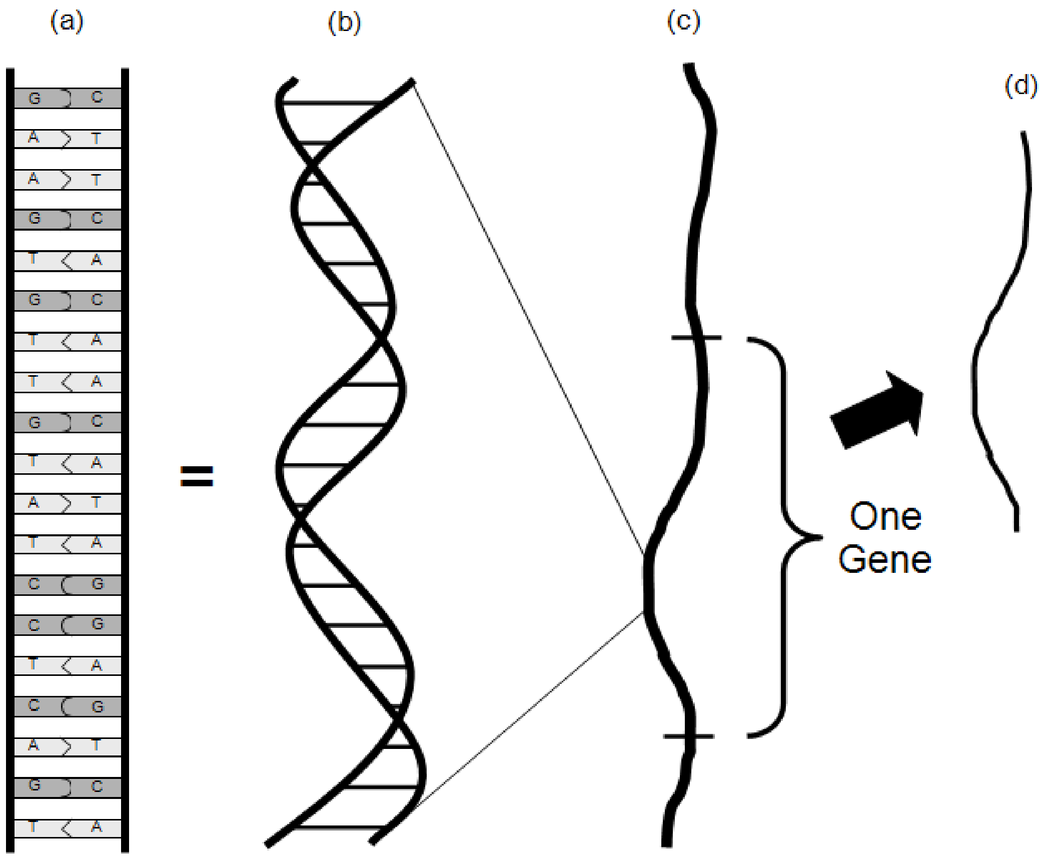
Examples of pre-processing .cel files using ***GESS*** are presented.

Overview

Gene Expression

The general process of gene expression in a cell begins with the DNA. Each DNA molecule has the well-known double helix design. Each rung or step of a DNA molecule is made up of two *nucleotides*. The two nucleotides bonded together are called *base pairs*. In DNA the 4 possible nucleotides are A, T, C, and G (for Adenine, Thymine, Cytosine, and Guanine, respectively). The nucleotide A can only form a base pair with T, and vice versa. Similarly, C can only bind to G, and vice versa. A gene is a unique segment of a DNA molecule consisting of a series of base pairs ranging from about 50 to thousands of base pairs in length. When the need for a specific protein in the cell is identified, the gene for that protein is "read" and a *messenger RNA* (mRNA) is produced in a process called *transcription*. mRNA molecules are single-stranded molecules which are essentially copies of the gene segment of one of the two DNA strands. The mRNA is then used to produce a protein, which is specific to that mRNA molecule, in a process called translation.

DNA Overview. (a) A ladder representation of a DNA segment showing 20 complementary base pairs. (b) A drawing of the three-dimensional form of the corresponding double helix. (c) The 20 base pairs of (a) and (b) are only a small section of the total DNA double strand. A gene is a segment of the DNA helix, which contains the code for the production of a protein. (d) A single stranded mRNA molecule is generated when the gene is expressed. The mRNA molecule will be used to produce a protein that is specific to the gene of (c).

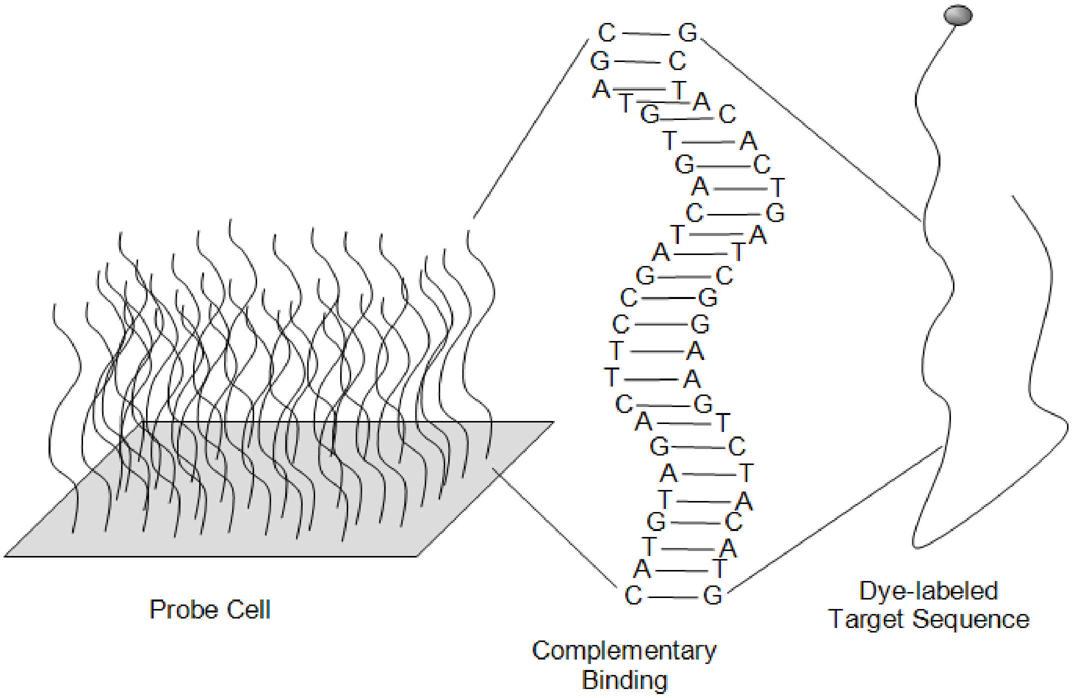


The newly created protein can then be used in the cell to perform the needed function. A gene that is in the process of producing, or has produced, a protein is said to be *expressed*. Expression of a given gene at a given time can thus be measured either by the amount of mRNA or protein (corresponding to that gene) in the cell. Microarrays are currently the prominent tool for quantifying the amount of mRNA in the cell (or collection of cells) for hundreds or thousands of genes simultaneously.

The Microarray

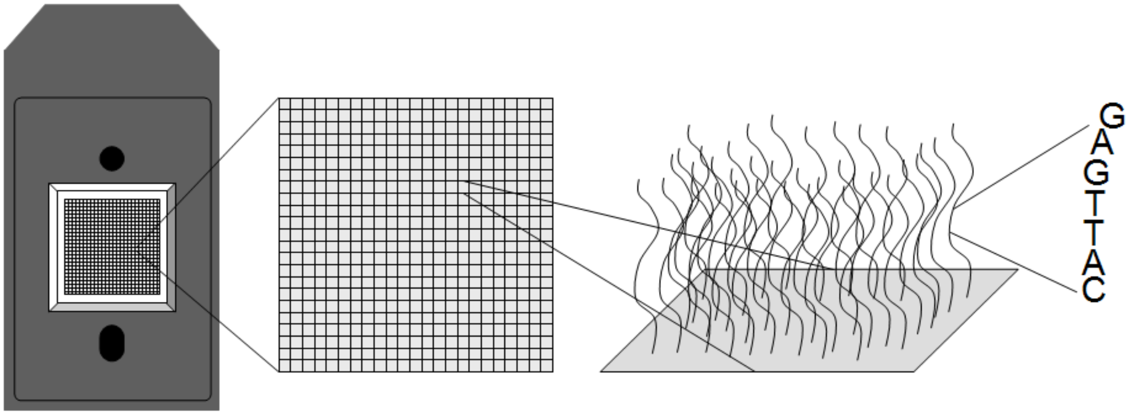
High-density oligonucleotide expression arrays are widely used in many areas of biology and medicine. Many competing arrays are on the market, but Affymetrix GeneChip® arrays are the most popular. A GeneChip® array is made up of a grid of hundreds of thousands of oligonucleotide *probes* (or probe cells). A probe is a tiny area (square) on the chip to which multiple copies of the same known sequence of 25 nucleotides has been attached. Those 25 nucleotides will bind with dye-labeled target sequences that complement those 25 nucleotides exactly.

Probe Cell. The 25-nucleotide region of the dye-labeled, single-stranded, gene expression mRNA sequence complements the 25-base oligonucleotide probe sequence. The probe will bind the target sequence. All sequences of the probe cell are identical.



Below is a microarray drawing depicting the arrangement of probe cells on a GeneChip® array (left), an enlarged view of a single section of the chip (left-center), the identical probe sequences immobilized on a single probe cell (right-center), and a segment of the probe for this cell that uniquely attracts cDNA strands of interest (right). The cDNA complement of the probe shown, from bottom to top, is GTAACTC.

GeneChip Drawing. A single chip consists of thousands of probe cells.



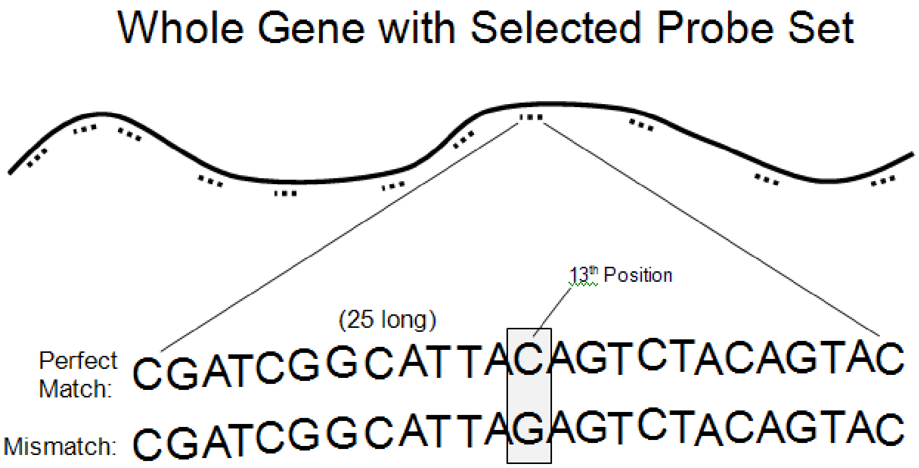
Array Design

Each gene (or genomic sequence) of interest is represented on the array by a probe set consisting of 11 to 20 probe pairs. Each pair is made up of a Perfect Match (PM) probe and a Mismatch (MM) probe. The sequences for the PM and MM probes are the same except for a single base substitution in the middle (13th position) of the MM probe sequence.

GeneChip® Array Design. Drawing of 11 probe pairs making up a probe set (shaded in gray) to measure expression of a single gene. On an actual Affymetrix GeneChip® array, there are tens of thousands of probe sets and hundreds of thousands of probe cells.

PM							PM			
MM							MM			
				PM					PM	
				MM					MM	
	PM									
	MM									
PM			PM					PM		
MM			MM					MM		
		PM		PM			PM			
		MM		MM			MM			

PM vs. MM Probes. Eleven 25-nucleotide probes (short dotted lines) are selected to represent the whole gene (solid line). Each of the 11 sequences is different from the other 10. One of the 11 sequences is enlarged and represented by its acronym. The mismatch sequence differs only at the 13th position.

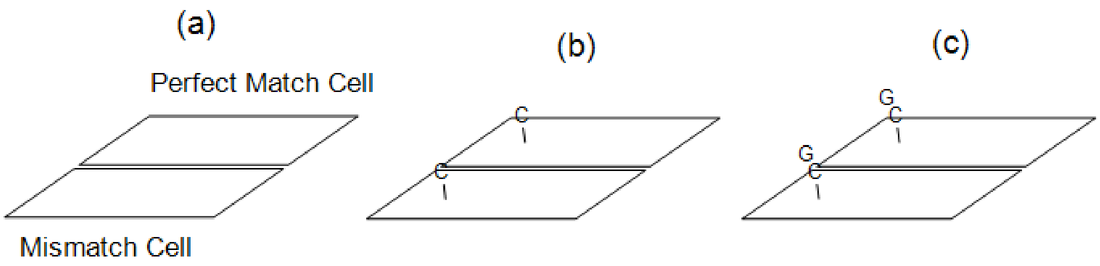

Whole Gene with Selected Probe Set
13th Position
(25 long)
Perfect Match:
CGATCGGCATTACAGTCTACAGTAC
Mismatch:
CGATCGGCATTAGAGTCTACAGTAC

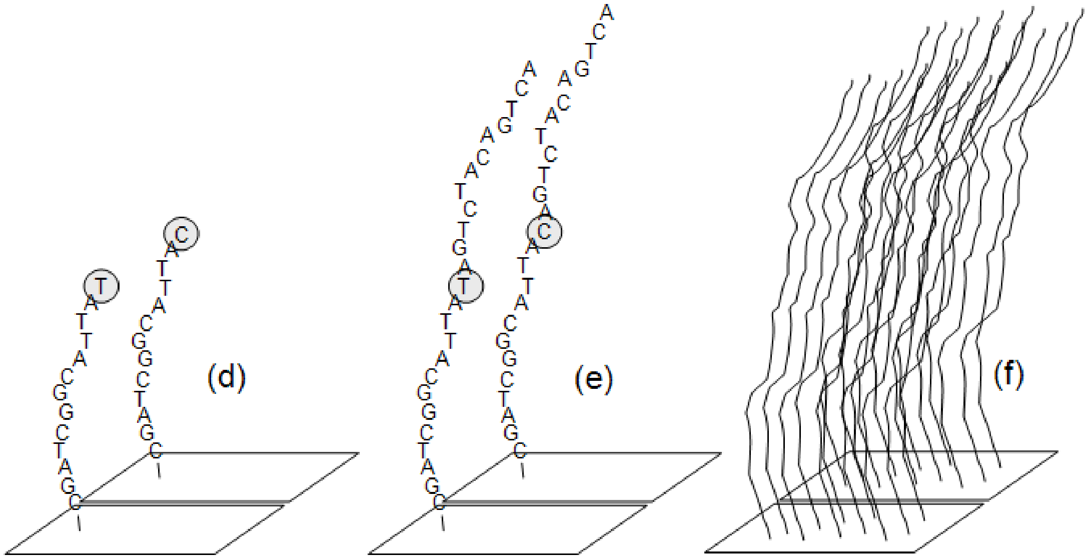
In Situ Fabrication

With *in situ* microarray synthesis, probe sequences are constructed nucleotide by nucleotide, directly on the chip. A general advantage of this method of probe synthesis is that the sequence of every probe is known exactly. A disadvantage is that *in situ* synthesis techniques limit the probe sequences to lengths much shorter than those of spotted probes.

Affymetrix GeneChip® arrays are created by attaching individual nucleotides one at a time to form highly specific probe sequences in a process called photolithography. Multiple copies of the same probe sequence are generated in a tiny square to make a probe cell. Perfect match and mismatch cell pairs are always adjacent on the array. Thousands of probe sets combine to make up all the cells in the array.

Array Fabrication. Synthesis of an Affymetrix GeneChip® probe shown for two cells in a cell pair. (a) Two cells make up a cell pair. (b) The first nucleotide is attached using photolithography. (c) The second nucleotide is attached. (d) The 13th nucleotide differs for the PM and MM sequences. (e) Completed probe sequences. (f) Each probe cell contains many identical sequences when the fabrication process is complete.





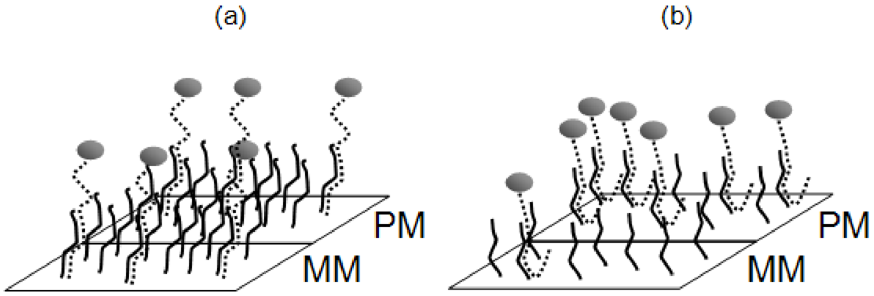
Five Steps to Obtain Relative Expression Values

Step 1 – Hybridization (Binding to the Microarray)

In an experiment, the mRNA expressed in an experimental unit is obtained from some of the experimental unit's cells (i.e., blood or tissue), converted to cDNA (an equivalent, but more stable molecule) using reverse transcription, and labeled with fluorescent dye. When the solution containing the cDNA is exposed to a GeneChip® array, the cDNA sequences will bind to the probe to which it complements. Thus, only sequences with perfect complementation along the 25-nucleotide region should bind to the corresponding probe. The cDNA from genes that are expressed in higher quantities will hybridize (bind) to the corresponding probe in higher quantities. The amount of hybridized material for each probe can then be measured using the intensity of dye fluorescence from the bound cDNA when exposed to laser scanning. The result is several thousand intensities that correspond in some degree to the amount of mRNA expression for each of those genes in that experimental unit.

One problem associated with microarray technology is *non-specific binding*. Non-specific binding refers to the case when a target sequence binds with a non-complementary probe. Each MM probe cell is designed to measure non-specific binding. The differing nucleotide at the 13th position should prevent the sample sequence from binding at that location. The MM probe, thus, was intended to provide a baseline or background measure. However, empirical evidence suggests that the MM probes may not be practically useful for measuring non-specific binding (see RMA Background Correction below).

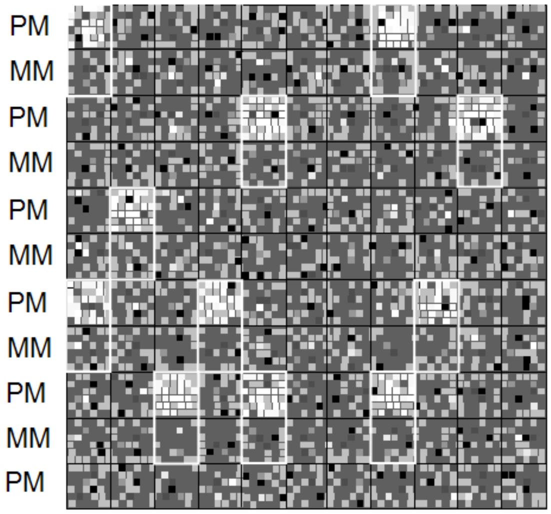
Hybridization on Adjacent PM/MM Probe Cells. (a) Hybridization at PM and MM cells is approximately equal, implying little or no signal at this location. (b) Hybridization at the PM cell is much higher than that of the MM cell, indicating the mRNA corresponding to this probe is likely highly expressed.



Step 2 – Image Construction

Following hybridization, a laser is used to illuminate the fluorescent dye across the array, producing an image of the thousands of cells. Each cell in the image is represented by a 6 × 6 = 36 pixel grid.

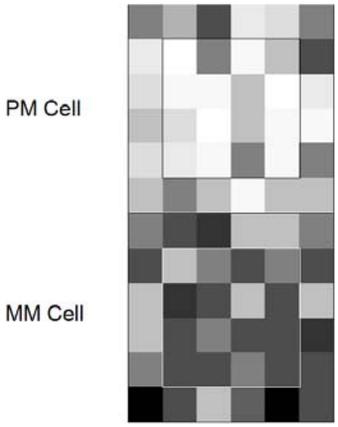
Drawing of a Segment of an Affymetrix GeneChip® Image. The probe pairs of probe set corresponding to a single gene are outlined in white. The PM probes of this probe set are much more illuminated than the MM probes, indicating evidence of expression of this gene.



Step 3 – Image Processing

The fluorescence of a single probe cell is summarized by determining the 75th percentile of the inner 16 pixels of that cell. These probe intensities are output by the scanner into a file with the extension ".cel".

Drawing of the Image for a Single GeneChip® Probe Pair. The 75th percentile of the middle 16 pixels is used as a measure of the intensity of each cell.



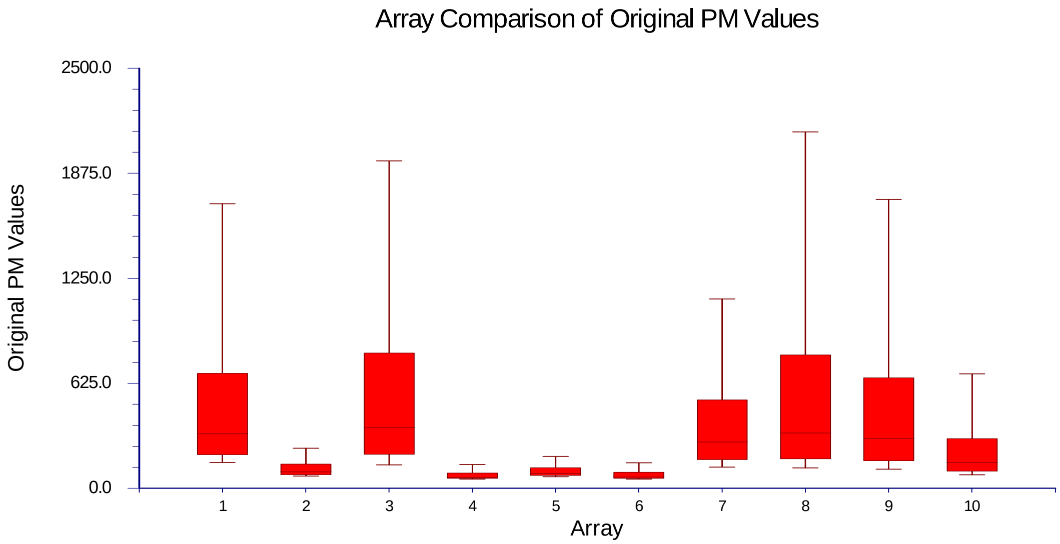
Step 4 – Whole-Array Quality Assessment

Much of the quality assessment of GeneChip® results is done using the proprietary techniques applied in associated Affymetrix software. Side-by-side box plots summarizing whole-chip expression values are useful for detecting chips with quality issues. Individual array intensity plots are also helpful for detecting abnormal spatial variation on an array.

Array Comparison Box Plot

All arrays in the experiment can be compared using a single side-by-side box plot. The box plot allows you to quickly compare the relative location and variation of chip PM intensities. The example below compares the original PM values from ten HG-U133A chips. It is instantly apparent from this plot that there is huge differences between medians and variation among chips.

Comparative Box Plot. 10 arrays are compared side-by-side on a single box plot.



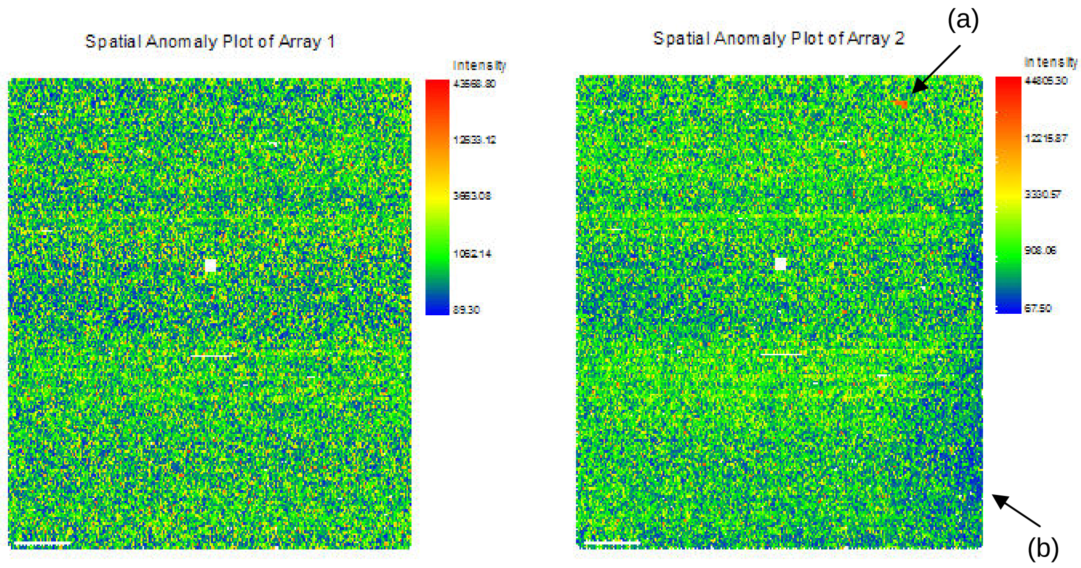
Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Spatial Anomaly Plots

A spatial anomaly plot is a reconstructed picture of the PM intensities. If there are no anomalies, there should be an even dispersion of low- and high-intensity probes throughout the plot.

The following are spatial anomaly plots of two HG-U133A arrays. A slight banding pattern from top to bottom is observed on both chips. Array #1 displays random scatter. Array #2 has a small area of high-intensity probes in the upper right hand corner that looks suspect. Array #2 also has a broad area of low-intensity probes in the bottom right quadrant. The vacant (white) areas on both chips correspond to the location of QC probes, where no PM intensities are found.:

Spatial Anomaly Plots. PM Intensity plots are used to look for spatial variation that indicates problematic chips. (a) A small region of non-random high-intensity probes. (b) A broad region of low-intensity probes.



Step 5 – Expression Index Calculation – RMA

In order to make accurate comparisons among chips, probe intensities are corrected for background noise, normalized, and summarized into a single expression measure for each gene from each array. Various methods have been proposed for background correction, normalization, and summarization of Affymetrix expression data. Of these, the Robust Multi-Array Average (RMA) algorithm has been shown to perform as well as or better than competing summarization algorithms in a variety of situations (see Bolstad et al. (2003) and Irizzary et al. (2003a,b)). The RMA algorithm consists of model-based background correction, quantile normalization, and median polish summarization as outlined in Irizzary et al. (2003a). We now treat each aspect of the RMA algorithm in detail.

RMA Background Correction

Microarray technology is designed to determine the expression level of a large number of genes simultaneously. In order to measure the true expression of a gene, it is important to separate the background signal, caused by non-specific binding and noise, from the true probe signal. The *MM* sequences were designed by Affymetrix to measure the background signal for each probe. The idea, in general, was to subtract the *MM* intensity from the *PM* intensity, resulting in an estimate of the intensity due to gene-specific complementary binding. However, results suggest that subtracting the *MM* intensities to correct for background noise is not always appropriate (Irizzary et al. (2003a), Naef et al. (2002)).

Irizzary et al. (2003a) presents a background correction method that uses *PM* values only (sometimes called RMA background correction). The method is based on a background plus signal model of the form

$$PM_{ijn} = bg_{ijn} + s_{ijn},$$

where $i = 1, 2, ..., I$ arrays, $j = 1, 2, ..., J_n$ probes, and $n = 1, 2, ..., N$ genes. The parameter bg_{ijn} represents the background signal, and s_{ijn} represents the true probe signal on array i. The *MM* intensities are not required in this algorithm. The true signal is estimated in the presence of background noise as

$$B(PM_{ijn}) \equiv E(s_{ijn} \mid PM_{ijn}).$$

If we impose a strictly positive distribution on s_{ijn}, this method forces all true signal estimates to be positive. Assume that

$$s_{ijn} \sim Exponential(\alpha_i)$$
$$bg_{ijn} \sim Normal(\mu_i, \sigma_i^{\,2}).$$

The parameters are estimated using only the *PM* intensity values on the array. Let β_i be the mode of the distribution of all *PM* intensities on the array. We estimate β_i as

$$\hat{\beta}_i = \text{the value of } PM_{ijn} \text{ at which the density trace}$$
$$\text{of all } PM \text{ intensities on array } i \text{ is maximized.}$$

The parameter β_i represents a minimum threshold expression level. Intensities below this level are assumed to be in the background range. We estimate μ_i from these background intensities as

$$\begin{aligned}\hat{\mu}_i = {} & \text{the value of } PM_{ijn} \text{ at which the density trace} \\ & \text{of } PM \text{ intensities less than } \hat{\beta}_i \text{ on array } i \text{ is maximized.}\end{aligned}$$

The standard deviation, σ_i, is estimated from the background intensities as

$$\hat{\sigma}_i = \sqrt{\frac{2\sum_{n=1}^{N}\sum_{j=1}^{J_n}\left(PM_{ijn}-\hat{\mu}_i\right)^2 I_{\{<\hat{\mu}_i\}}\left(PM_{ijn}\right)}{\sum_{n=1}^{N}\sum_{j=1}^{J_n} I_{\{<\hat{\mu}_i\}}\left(PM_{ijn}\right)-1}}\ .$$

To estimate the exponential parameter, α_i, we first select the subset, PM_i*, of *PM* intensities from array *i* that are greater than $\hat{\mu}_i$,

$$PM_i{}^* = \{PM_{ijn} > \hat{\mu}_i\}\ .$$

Let *k* index the members of PM_i*. We now subtract $\hat{\mu}_i$ from each member of PM_i* to obtain the transformed data subset PM_i',

$$PM_i{}' = PM_i{}^* - \hat{\mu}_i\ .$$

Let η_i be the mode of the distribution of all PM_i' intensities. We estimate η_i as

$$\begin{aligned}\hat{\eta}_i = {} & \text{the value of } PM_{ik} \text{ at which the density trace} \\ & \text{of all } PM_i{}' \text{ intensities is maximized.}\end{aligned}$$

Finally, estimate α_i as

$$\hat{\alpha}_i = \frac{1}{\hat{\eta}_i}\ .$$

We then background correct the *PM* values for array *i* as (Wit an McClure (2004) pages 78 and 79)

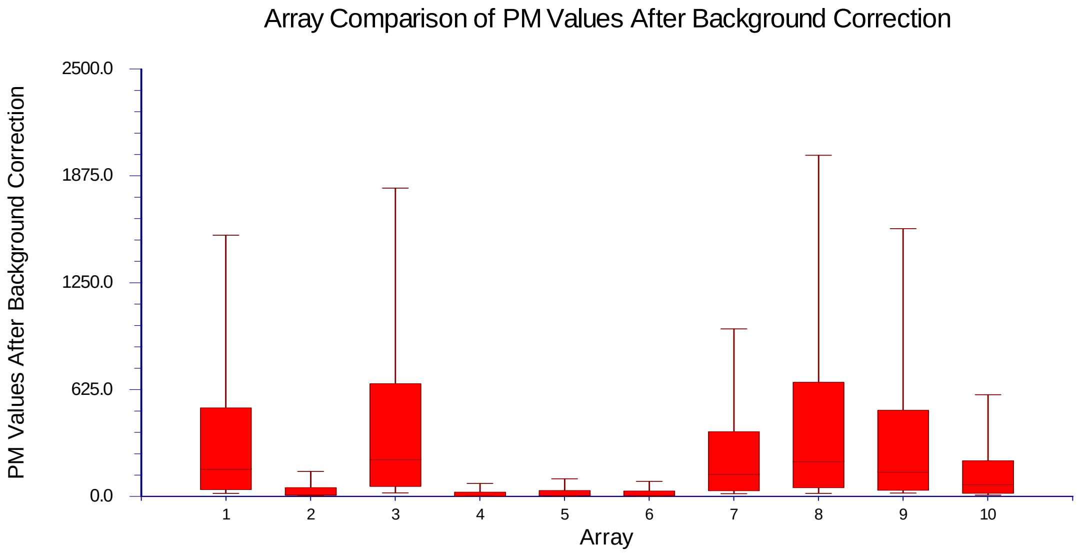
$$\begin{aligned}B\left(PM_{ijn}\right) \equiv E\left(s_{ijn} \mid PM_{ijn}\right) &\cong PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i{}^2\hat{\alpha}_i + \hat{\sigma}_i \frac{\varphi\left(\frac{PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right) - \varphi\left(\frac{\hat{\mu}_i + \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right)}{\Phi\left(\frac{PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right) + \Phi\left(\frac{\hat{\mu}_i + \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right) - 1} \\ &= PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i{}^2\hat{\alpha}_i + \hat{\sigma}_i \frac{\varphi\left(\frac{PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right) - \varphi\left(\frac{\hat{\mu}_i + \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right)}{\Phi\left(\frac{PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right) + \Phi\left(\frac{\hat{\mu}_i + \hat{\sigma}_i{}^2\hat{\alpha}_i}{\hat{\sigma}_i}\right) - 1}\end{aligned}$$

$$\approx PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i^{\,2}\hat{\alpha}_i + \hat{\sigma}_i \frac{\varphi\left(\dfrac{PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i^{\,2}\hat{\alpha}_i}{\hat{\sigma}_i}\right)}{\Phi\left(\dfrac{PM_{ijn} - \hat{\mu}_i - \hat{\sigma}_i^{\,2}\hat{\alpha}_i}{\hat{\sigma}_i}\right)}$$

where φ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively.

From the plot below, we can see that the background correction algorithm applied to the same ten chips plotted above (see Array Comparison Box Plot above) moves the boxes nearer to zero. The resulting PM values no longer contain a background signal.

Comparative Box Plot. PM values after background correction.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Quantile Normalization

The need for normalization arises naturally when multiple arrays are used in an experiment. Variation arising from small differences in sample preparation, array production, and scanner processing may obscure the interesting differences in gene expression due to treatment or sample group. The goal of quantile normalization is to make the distribution of *PM* probe intensities the same for each array in a set of arrays, allowing for accurate comparison between arrays.

Quantile normalization in ***GESS*** is carried out according to the algorithm described in Bolstad et al. (2003). The algorithm is presented below. An example data matrix is provided for illustration.

1. Given *n* arrays, each containing *p PM* probe intensities, form *X* of dimension *p* x *n* where each array is a column and each *PM* location corresponds to a row.

$$\begin{array}{cc} & \text{Array} \\ & \begin{array}{cccc}\mathbf{1}&\mathbf{2}&\mathbf{3}&\mathbf{4}\end{array} \\ \text{PM}\quad \begin{array}{c}\mathbf{1}\\\mathbf{2}\\\mathbf{3}\\\mathbf{4}\\\mathbf{5}\end{array} & \begin{pmatrix}7&8&9&8\\4&2&1&2\\9&8&2&2\\8&7&6&2\\1&7&3&5\end{pmatrix} = X\end{array}$$

2. Sort each column of *X* individually from largest to smallest to give X_{sort}. Rank each element of *X* within columns to get X_{rank}. Fractional ranks are rounded to the nearest whole.

$$X_{sort} = \begin{pmatrix}9&8&9&8\\8&8&6&5\\7&7&3&2\\4&7&2&2\\1&2&1&2\end{pmatrix} \qquad X_{rank} = \begin{pmatrix}3&2&1&1\\4&5&5&4\\1&2&4&4\\2&4&2&4\\5&4&3&2\end{pmatrix}$$

3. Take the means across each row of X_{sort} to get a vector of mean intensities. This provides a mapping vector for translating ranks into normalized intensity values. In our example, a rank of one corresponds to 8.5, a rank of two corresponds to 6.75, and so forth.

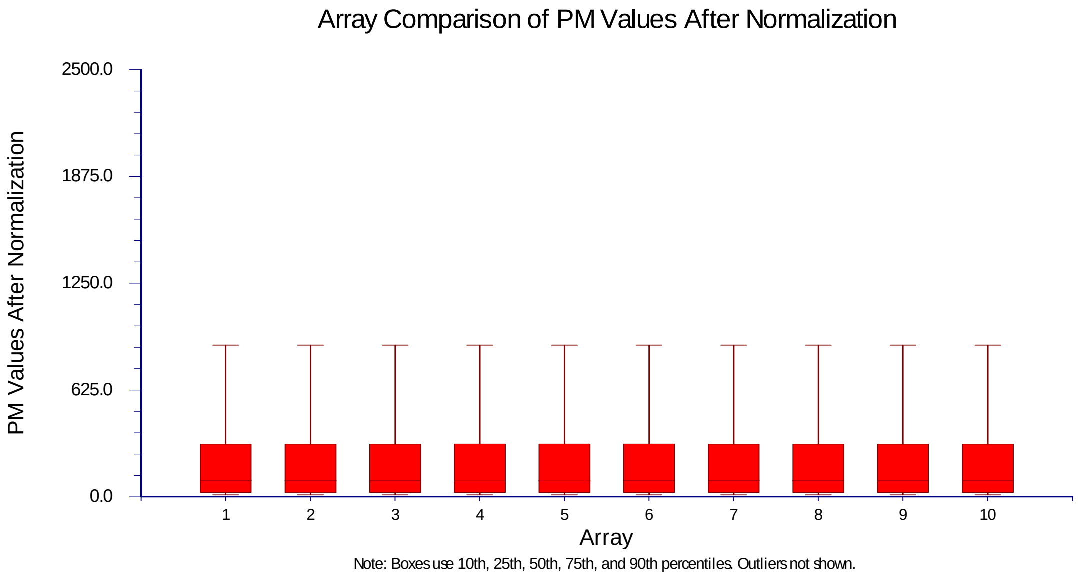
$$X_{sort} = \begin{pmatrix}9&8&9&8\\8&8&6&5\\7&7&3&2\\4&7&2&2\\1&2&1&2\end{pmatrix} \begin{matrix}\xrightarrow{\qquad\qquad\qquad\qquad}\\ \xrightarrow{\qquad\qquad\qquad\qquad}\\ \text{Mean Mapping Vector} = \\ \xrightarrow{\qquad\qquad\qquad\qquad}\\ \xrightarrow{\qquad\qquad\qquad\qquad}\end{matrix} \begin{pmatrix}8.5\\6.75\\4.75\\3.75\\1.5\end{pmatrix} \quad \begin{array}{c}\text{Rank}\\ \mathbf{1}\\\mathbf{2}\\\mathbf{3}\\\mathbf{4}\\\mathbf{5}\end{array}$$

4. Get $X_{Normalized}$ by replacing each element in X_{rank} with the corresponding mean from the mean mapping vector.

$$X_{rank} = \begin{pmatrix}3&2&1&1\\4&5&5&4\\1&2&4&4\\2&4&2&4\\5&4&3&2\end{pmatrix} \qquad X_{Normalized} = \begin{pmatrix}4.75&6.75&8.5&8.5\\3.75&1.5&1.5&3.75\\8.5&6.75&3.75&3.75\\6.75&3.75&6.75&3.75\\1.5&3.75&4.75&6.75\end{pmatrix}$$

The plot below demonstrates that even though there was a huge difference in variation and medians between chips (see Array Comparison Box Plot above), the quantile normalization algorithm creates equal distributions of PM values for all arrays. These arrays can now be compared to each other.

Comparative Box Plot. PM values after background correction and normalization.



Summarization - Median Polish

GESS implements the Robust Multi-Array Average (RMA) algorithm for expression index calculation as presented in Irizarry et al. (2003b). The RMA algorithm is based on an additive model of the form

$$T(PM_{ijn}) = e_{in} + a_{jn} + \varepsilon_{ijn} \,,$$

where $i = 1, 2, ..., I$ arrays, $j = 1, 2, ..., J_n$ probes, and $n = 1, 2, ..., N$ genes. The term $T(\bullet)$ represents the transformation that background corrects, normalizes, and takes the $\log_2$ of the *PM* intensity values. The factor e_{in} represents the $\log_2$ scale expression value for gene *n* found on array *i*. The factor a_{jn} represents the log-scale affinity effect for gene *n* and probe *j*. The factor ε_{ijn} represents the independent and identically distributed error with mean 0. For identifiability of the parameters, we assume that for all probe sets

$$\sum_j a_{jn} = 0 \,.$$

Median Polish is used to estimate the model parameters robustly. The median polish algorithm is described in Tukey (1977) and proceeds as follows. A simplified example is provided for illustration.

1. Form a J_n x I matrix of $\log_2$-scale *PM* intensity values for gene *n* that have been background corrected and normalized (if desired).

$$
\begin{array}{cc|cccc|}
 & & \multicolumn{4}{c}{\text{Array}} \\
 & & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\
 & \mathbf{1} & 9 & 12 & 10 & 7 \\
 & \mathbf{2} & 8 & 9 & 10 & 7 \\
 & \mathbf{3} & 9 & 10 & 10 & 12 \\
 & \mathbf{4} & 12 & 9 & 12 & 12 \\
 & \mathbf{5} & 12 & 8 & 12 & 10 \\
\text{Probe} & \mathbf{6} & 10 & 9 & 11 & 11 \\
 & \mathbf{7} & 7 & 8 & 9 & 9 \\
 & \mathbf{8} & 8 & 10 & 9 & 10 \\
 & \mathbf{9} & 8 & 7 & 8 & 12 \\
 & \mathbf{10} & 9 & 12 & 8 & 9 \\
 & \mathbf{11} & 10 & 8 & 9 & 8
\end{array}
$$

2. Proceed by iteratively subtracting off the column median from each value in the corresponding column followed by the row median from each value in the corresponding row. Stop when elements of the matrix no longer change by a significant amount. The resulting matrix contains the residuals of the median polish fit.

$$
\begin{array}{cc|cccc|}
 & & \multicolumn{4}{c}{\text{Array}} \\
 & & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\
 & \mathbf{1} & 0.00 & 3.00 & 0.00 & -2.75 \\
 & \mathbf{2} & -0.50 & 0.50 & 0.50 & -2.75 \\
 & \mathbf{3} & -0.50 & 0.50 & -0.50 & 1.75 \\
 & \mathbf{4} & 0.75 & -2.25 & -0.25 & 0.00 \\
 & \mathbf{5} & 1.75 & -2.25 & 0.75 & -1.00 \\
\text{Probe} & \mathbf{6} & 0.00 & -1.00 & 0.00 & 0.25 \\
 & \mathbf{7} & -1.00 & 0.00 & 0.00 & 0.25 \\
 & \mathbf{8} & -0.75 & 1.25 & -0.75 & 0.50 \\
 & \mathbf{9} & 0.50 & -0.50 & -0.50 & 3.75 \\
 & \mathbf{10} & 0.25 & 3.25 & -1.75 & -0.50 \\
 & \mathbf{11} & 2.00 & 0.00 & 0.00 & -0.75
\end{array}
$$

3. Subtract the residual matrix from the original data matrix to obtain the fitted values. Find the median of each column of the fitted value matrix to obtain the final expression estimates for each array. In our example, the final expression estimates are 8.75, 8.75, 9.75, and 9.50 for gene *n* on arrays 1 through 4, respectively.

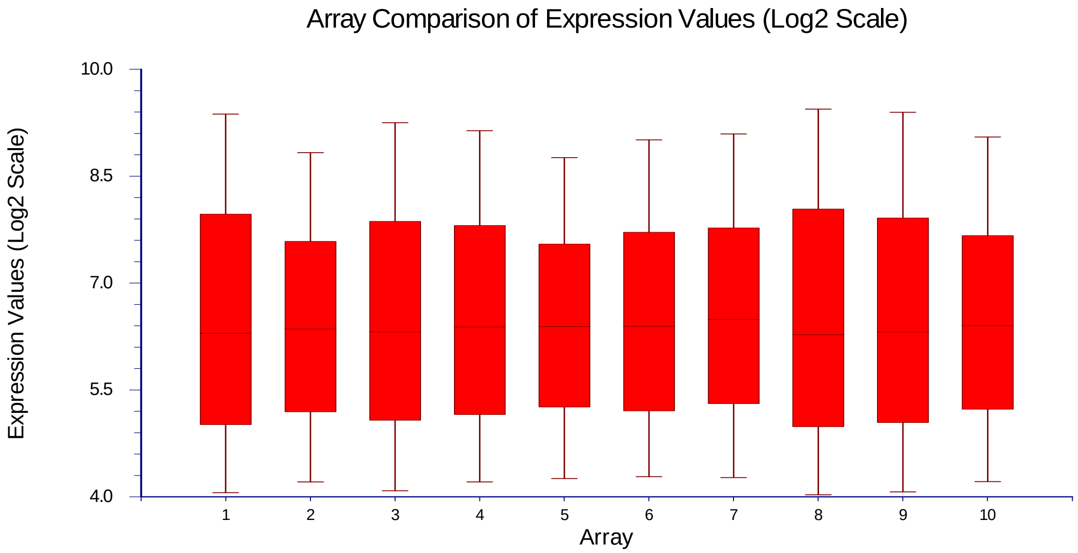
$$\begin{pmatrix} 9 & 12 & 10 & 7 \\ 8 & 9 & 10 & 7 \\ 9 & 10 & 10 & 12 \\ 12 & 9 & 12 & 12 \\ 12 & 8 & 12 & 10 \\ 10 & 9 & 11 & 11 \\ 7 & 8 & 9 & 9 \\ 8 & 10 & 9 & 10 \\ 8 & 7 & 8 & 12 \\ 9 & 12 & 8 & 9 \\ 10 & 8 & 9 & 8 \end{pmatrix} - \begin{pmatrix} 0.00 & 3.00 & 0.00 & -2.75 \\ -0.50 & 0.50 & 0.50 & -2.75 \\ -0.50 & 0.50 & -0.50 & 1.75 \\ 0.75 & -2.25 & -0.25 & 0.00 \\ 1.75 & -2.25 & 0.75 & -1.00 \\ 0.00 & -1.00 & 0.00 & 0.25 \\ -1.00 & 0.00 & 0.00 & 0.25 \\ -0.75 & 1.25 & -0.75 & 0.50 \\ 0.50 & -0.50 & -0.50 & 3.75 \\ 0.25 & 3.25 & -1.75 & -0.50 \\ 2.00 & 0.00 & 0.00 & -0.75 \end{pmatrix} = \begin{array}{c} \text{Array} \\ \begin{array}{cccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \end{array} \\ \begin{pmatrix} 9.00 & 9.00 & 10.00 & 9.75 \\ 8.50 & 8.50 & 9.50 & 9.25 \\ 9.50 & 9.50 & 10.50 & 10.25 \\ 11.25 & 11.25 & 12.25 & 12.00 \\ 10.25 & 10.25 & 11.25 & 11.00 \\ 10.00 & 10.00 & 11.00 & 10.75 \\ 8.00 & 8.00 & 9.00 & 8.75 \\ 8.75 & 8.75 & 9.75 & 9.50 \\ 7.50 & 7.50 & 8.50 & 8.25 \\ 8.75 & 8.75 & 9.75 & 9.50 \\ 8.00 & 8.00 & 9.00 & 8.75 \end{pmatrix} \\ \begin{array}{cccc} \mathbf{8.75} & \mathbf{8.75} & \mathbf{9.75} & \mathbf{9.50} \end{array} \end{array}$$

4. Repeat steps 1 through 3 for each gene represented on the arrays to obtain expression estimates for all the genes. The final expression estimates are on the $\log_2$ scale. They may be transformed to any scale.

The result of median polish is an expression value for each probe set from each chip. These expression values can now be compared with each other in further statistical analyses.

The plot below shows the relative expression values on the Log_2 scale for the ten chips after summarization by median polish. The medians are constant, and the variation is roughly equal for all ten chips. These chips can now be compared in further statistical analyses.

Comparative Box Plot. Expression measures on the Log_2 scale after background correction, normalization, and summarization by median polish.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Affymetrix File Types

Several different files are generated and used by the Affymetrix GeneChip® array system. Of these files, ***GESS*** requires only two for complete Affymetrix microarray analysis: the .cdf library file and the .cel intensity files. The .cdf file contains the information necessary to interpret the data in the .cel file. The .cel files contain the summarized intensity information for each probe. Both files are required in order to associate intensity readings with specific probes.

CDF Files

The .cdf file contains the array "map" necessary for interpretation of the .cel files in an experiment. The .cdf file outlines probe set membership and indicates the location of PM and MM probes in each probe set. The .cdf file also contains information about quality control probes necessary for scanner alignment and array quality assessment. ***GESS*** does not use any of the quality control probe information. The .cdf file contains the following information about experimental probes.

X	The x coordinate of the probe cell on the chip.
Y	The y coordinate of the probe cell on the chip.
PROBE	The probe sequence of the cell. Typically set to "N".
QUAL	The same as the block name, i.e. the name of the probe set.
EXPOS	Ranges from 0 to the number of atoms - 1 for expression arrays. For Customseq it provides some positional information for the probe.
POS	An index to the base position within the probe where the mismatch occurs.
PBASE	The probe base at the substitution position.
TBASE	The base of the target where the probe interrogates at the substitution position.
INDEX	An index for the corresponding probe cell data in the .cel file.

CEL Files

The .cel files contain the fluorescence intensity data generated by reading an Affymetrix GeneChip® array. The .cel file consists of several subsections, including a header with information about the chip setup, analysis parameters, and intensity data. Affymetrix .cel files exist in both binary and text file formats. ***GESS*** can read either format. Each .cel file contains the following information about every individual probe.

X	The x coordinate of the cell on the chip.
Y	The y coordinate of the cell on the chip.
MEAN	The probe cell intensity value. This is usually the 75th percentile of pixel intensities observed for a given probe. The number of pixels used in this calculation is given by NPIXELS.
STDV	The standard deviation of the pixel intensities.
NPIXELS	The number of pixels used for intensity determination at a single probe site.

The .cel files also contain information about masked probes and outlier probes.

Entering CDF and CEL Files

This section describes how file names are entered into the spreadsheet in preparation for pre-processing. Two variables (columns) are required to run Affymetrix pre-processing, and a third is required to obtain output files. Any name (entered by clicking on the Variable Info tab at the bottom of the spreadsheet) may be used for these variables.

CDF File Name Variable

The CDF File Name Variable is a column on the spreadsheet containing the filename and path of the .cdf file that corresponds to the chips from which the .cel files were created. The name and path must be entered in the **first cell** of this column. This variable is required to run the Affymetrix CEL File Pre-Processing Engine.

To enter the .cdf file name and path into the spreadsheet:

1. Highlight the first cell in an empty column.
2. Either type in the path and file name directly or hit **F7** to browse for the appropriate .cdf file.

CEL File Names Variable

The CEL File Names Variable is a column on the spreadsheet containing a list of filenames and paths of the .cel files that are to be included in pre-processing. The files may be in different folders but all must correspond to the same array type (.cdf file). This variable is required to run the Affymetrix Pre-Processing Engine.

To enter a .cel file name and path into the spreadsheet:

1. Highlight an empty cell.
2. Either type in the path and file name directly or hit **F7** to browse for an appropriate .cel file.
3. Repeat steps 1 and 2 until all .cel files have been entered.

Output File Names Variable

When Affymetrix pre-processing is run, a new set of files is generated automatically for use in statistical analyses. These output files have the extension .ges, and are stored in the folder specified under Folder in which Output Files will be Stored. The path and name of each .ges output file is placed on the spreadsheet in the column specified by the Output File Names Variable. This column of pre-processed output files will become your input files for further statistical analyses. This variable is required to obtain output for statistical analysis. If this variable is left blank, no new .ges output files will be created.

To specify an output folder under Folder in which Output Files will be Stored:

1. Click on the **Browse** button to the right of the window.
2. Select an output folder and click **OK**.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options specify the variables and major pre-processing options that will be used in the analysis.

Affymetrix Files for Pre-Processing

These options are used specify the input .cdf file and the .cel files that are to be pre-processed.

CDF File Name Variable

Select the variable that contains the .cdf file for the chip used in the experiment.

The name and path of the file must appear in the **first cell** of the column below this variable name on the spreadsheet. To enter the .cdf file name and path into the spreadsheet:

1. Highlight the first cell in an empty column.
2. Either type in the path and file name directly or hit F7 to browse for the appropriate .cdf file.

When a .cdf file is first used in pre-processing, ***GESS*** will automatically create and store a new file containing the .cdf file information. This new file is stored in the *C:\...\[My] Documents \NCSS\NCSS 2007\Data\GESS\CDH* folder and has the extension ".cdh". ***GESS*** reads the .cdh file much faster than it reads the corresponding .cdf file. In subsequent runs of the Affymetrix Pre-Processing Engine, ***GESS*** will preferentially use the stored .cdh file to increase efficiency.

A new .cdh file is stored for each different .cdf file used. For example, if you processed .cel files using HG-U133A.cdf, a new file with the name "HG-U133A.cdh" would be created. On subsequent runs using HG-133A chips, the HG-U133A.cdh file will be used. If at some point you processed new .cel files using HG-Focus.cdf, the file "HG-Focus.cdh" would be stored. If you believe that a .cdf file you are using has changed since you last ran the engine, you should delete the current .cdh file from the *C:\...\[My] Documents\NCSS\NCSS 2007\Data\GESS\CDH* folder before running the procedure.

CEL File Names Variable

Select the variable that contains the list of the .cel files from the experiment.

The names and pathways of the files should appear in a column below this variable name on the spreadsheet. To enter the .cel file names and paths into the spreadsheet:

1. Highlight an empty cell.
2. Either type in the path and file name directly or hit **F7** to browse for an appropriate .cel file.
3. Repeat steps 1 and 2 until all .cel files have been entered.

RMA Pre-Processing Options

These options determine the summary value that will be produced in the output files.

Background Correction

Check this box to perform RMA (Model-Based) background correction. If this box is left unchecked, no background correction will be performed.

Background correction is performed prior to both normalization and median polish expression summarization. The goal of background correction is to separate the true probe signal from the background signal caused by non-specific probe binding and noise.

Quantile Normalization

Check this box to perform Quantile Normalization. If this box is left unchecked, no normalization will be performed.

Normalization is performed after background correction but before median polish expression summarization. The need for normalization arises naturally when multiple arrays are used in an experiment. The goal of quantile normalization is to make the distribution of PM probe intensities the same for each array in a set of arrays, allowing for accurate comparison between arrays. If normalization is not performed, interesting differences in expression may be obscured by variation arising from small differences in sample preparation, array production, and scanner processing.

Expression Measure Output Scale

This option allows you to select the scale for the final gene expression estimates output by RMA Median Polish. The options are Original Scale, Log_2, Log_e (natural log), and Log_{10}.

Median polish is a robust algorithm for calculating expression measures for oligonucleotide probe-level data. Median polish is performed after both background correction and normalization have been completed according to the algorithm presented in Irizarry et al. (2003b). Prior to median polish summarization, the intensity estimates are transformed to the $\log_2$ scale.

GES Output Files Specifications

These options are used to specify the location and naming of the output .ges files.

Folder in which Output Files will be Stored

Enter the path and name of the folder in which the newly created .ges output files will be stored. The path may be typed directly, or the Browse button may be used to locate the desired folder.

In the default path "%mydocs_NCSS%\DATA\GESS", the designator "%mydocs_NCSS%" represents the path to the ***GESS*** personal data folder (commonly *C:\...\[My] Documents\NCSS \NCSS 2007*).

Output File Names Variable

Select the variable in which the path and names of the newly created .ges files will be stored for future statistical analyses. The path and folder for these .ges files is entered under Folder in which Output Files will be Stored.

A new file (to be used for statistical analyses) will be created for each row when the procedure is run.

If this variable is left blank, no new .ges files will be created.

Append to File Names

A sequence of characters may be entered here to append to the name of the created file.

For example, if the .cel file has the name "Slide1_10hours.cel" and "log" is entered here, the newly created .ges file will be "Slide1_10hours log.ges".

If nothing is entered here, the file name will be the same as the name of the .cel file, but ".cel" will be replaced with ".ges".

Overwrite existing output (.ges) files with new output (.ges) files

Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file.

For example, if "Slide1.ges" has already been created and a new "Slide1.ges" file is to be written, the new file will be "Slide1 (2).ges" if the Overwrite box is not checked.

Other Options

These options determine the summary value that will be produced in the output files.

Generate Quality Control Reports and Plots Only

Check this box to generate the PM Intensity Summary Report, Spatial Anomaly Plots, and the Comparative PM Intensity Box Plot without proceeding to background correction, normalization, and median polish. If this box is checked, no output files will be stored.

This option allows you to look for individual array problems before proceeding to the somewhat lengthy process of computing expression estimates.

NOTE: When you check this option, you must still select the desired reports on the Reports tab.

Reports Tab

The options on this panel control which reports and plots are generated.

Select Reports

The following reports and report options are available.

File Processing Summary

Check this box to obtain a row-by-row summary of input and output file names and a summary of pre-processing specifications.

Data Summary

Check this box to obtain a row-by-row summary of PM intensities and expression values for each array.

Decimals

Specify the number of decimals to be used for percentiles in the Data Summary report. The number of decimal places is used for output display only. It does not change the internal precision of the data.

Select Plots

Choose from the following plots.

Comparative Box Plots

Check this box to output comparative box plots. These box plots allow you to compare the original PM intensities, the background-corrected PM intensities, the normalized PM intensities, and the expression values from each array. The settings for the comparative box plots are specified under the Box Plot tab.

Spatial Anomaly Plots

Check this box to output the spatial anomaly plot for each array.

The spatial anomaly plot gives a spatial view of the PM intensities for the entire array. The settings of the spatial anomaly plot are specified under the Spatial Plot tab.

NOTE: Checking this box will cause a notable increase in computing time and output display (refresh) time for large datasets.

Options Tab

This panel contains miscellaneous options.

Median Polish Options

The following option is available to control median polish processing.

Chunk Size

This option allows you to specify how many probe sets are read into memory during median polish. Increasing this number will speed up the process but may cause you to run out of memory for very large numbers of arrays. If you find that you are running out of memory during median polish, decrease this number.

Box Plots Tab

The options on this panel control attributes of the box plots.

Horizontal and Vertical Axes

The following options allow you to format the horizontal (X) and vertical (Y) axes.

Label

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate X-axis label.

{Y} is replaced by an appropriate Y-axis label.

Ref. Number Format...

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Major Ticks

Specify the number of large tickmarks and optional grid lines along the axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along the axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along the axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the corresponding axis. Thus, checking Show Grid Lines here will actually cause horizontal grid lines to appear.

Vertical Minimums and Maximums

The following options allow you to set the vertical (Y) axis minimum and maximum separately for the PM Intensity and Expression Value box plots.

Minimum

Specify the value to be displayed as the minimum on this axis. If this value is left blank, the minimum will be determined from the data. If this value is greater than the smallest 10th percentile of the data, it will be ignored.

Maximum

Specify the value to be displayed as the maximum on this axis. If this value is left blank, the maximum will be determined from the data. If this value is less than the largest 90th percentile of the data, it will be ignored.

Box Plot Settings

The following options allow you to control the appearance of the box plot.

Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created in the Box Plots procedure.

% Space

When the Box Width (or Bar Width) option is set to Percent Space in the box plot style file selected, this value specifies the percent of the length of the axis that is empty space instead of

bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also, note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**

  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**

  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Interior

The color used to fill the rectangle formed by the vertical and horizontal axes. Click to change.

Background

The color used behind the plot. Click to change.

Box Fill

The color used to fill the boxes. Click to change.

Box Border

The color used to outline the boxes. Click to change.

Line

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Top and Bottom Titles

Enter text here for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate horizontal-axis label.

{Y} is replaced by an appropriate vertical-axis label.

Spatial Anomaly Plots Tab

The options on this panel control the features of the spatial anomaly plots.

Heat Map Settings

The following options allow you to control the heat map settings.

Heat Map Colors and Scale

Click on the heat map color bar or the button to the right to change the colors and/or scale of the heat map. The scaling options are Regular, Percentile, and Log. By default, the log scale is used.

Probes Skipped

Some microarrays have so many probes that the plot is difficult to display rapidly. This option lets you cut down on the refresh time by allowing you to skip a specific number of rows and columns displayed on the plot. If you select "Automatic", ***GESS*** will optimize the number of rows to skip for the size of array being plotted.

Legend

Use the following options to control the display of the heat map legend.

Label

Enter text here for the label of the heat map legend. Click the arrow to the right of the box to modify the label format.

Number of Values

This is the number of reference values printed along the right side of the heat map legend.

Show Legend

Specify whether to show the legend.

Value Format

This option specifies the characteristics of the reference numbers shown next to the heat map legend.

It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot.

It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Plot Settings

The following options allow you to control the spatial anomaly plot settings.

Style File

A plot style file sets all plot options that are not set directly by this procedure.

Interior

Specify the interior color of the spatial anomaly plot.

Background

Specify the background color of the spatial anomaly plot.

Titles

Enter text here for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by the name of the appropriate .cel file.

{Y} is replaced by the name of the GeneChip array (e.g. "HG-U133A").

{Z} is replaced by the appropriate array row number.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Pre-Processing Affymetrix CEL Files

This section presents an example of how to pre-process six .cel files. The spreadsheet data used are recorded in the AFFYCEL dataset. The input files are stored in the *C:\Program Files\NCSS \NCSS 2007\Data\GESS\AF* directory.

To run this example, take the following steps or load the **Example 1** template from the Affymetrix CEL File Pre-Processing Engine Template tab.

1 Open the AFFYCEL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **AFFYCEL.S0**.
- Click **Open**.

2 Open the Affymetrix CEL File Pre-Processing Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Affymetrix CEL Files**. The Affymetrix CEL File Pre-Processing Engine procedure will be displayed.
- On the Affymetrix CEL File Pre-Processing Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the Variables.

- On the Affymetrix CEL File Pre-Processing Engine window, select the **Variables tab**.
- Set the **CDF File Name Variable** to **CDFFile**.
- Set the **CEL File Names Variable** to **InputFile**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **OutputFile**.
- Check the box next to **Overwrite existing output (.ges) files with new output (.ges) files**.
- Leave all other options under the Variables tab and other tabs at their default settings.

4 Specify the reports.

- Select the **Reports tab**.
- Check the box next to **Spatial Anomaly Plots**.
- Leave all other options under the Reports tab at their default settings.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Input File Summary

```
Input File Summary

         Number of
Row      Probes        File Name
1        15876         ...\DATA\GESS\AF\AffyCEL_Ex1_1.cel
2        15876         ...\DATA\GESS\AF\AffyCEL_Ex1_2.cel
3        15876         ...\DATA\GESS\AF\AffyCEL_Ex1_3.cel
4        15876         ...\DATA\GESS\AF\AffyCEL_Ex1_4.cel
5        15876         ...\DATA\GESS\AF\AffyCEL_Ex1_5.cel
6        15876         ...\DATA\GESS\AF\AffyCEL_Ex1_6.cel

CDF File Name: ...\DATA\GESS\AF\Test3.cdf
CDH File Name: ...\DATA\GESS\CDH\Test3.cdh
```

This report displays a list of input (.cel) files and the number of probes (PM, MM, and Affymetrix QC) encountered in each file. Only PM values are used in plots and pre-processing. This report also gives the names of the .cdf file entered and the corresponding .cdh file (see CDF File Name Variable above for a description of the .cdh file).

Row

This is the row of the array on the spreadsheet.

Number of Probes

This contains the total number of probes encountered in each .cel file.

File Name

This contains the path and name of each input (.cel) file.

Pre-Processing Methods Summary

```
Pre-Processing Methods Section

Task Name                 Method Selected
Background Correction     RMA (Model-Based)
Normalization             Quantile
Summarization             Median Polish
Output Scale              Log base 2
```

This report shows the pre-processing methods selected.

Task Name

These are the different tasks that may be completed during RMA expression summarization.

Method Selected

These are the pre-processing methods selected for each task.

Output File Summary

```
Output File Summary

        Number of
Row     Probe Sets    File Name
1       345           ...\DATA\GESS\AffyCEL_Ex1_1.ges
2       345           ...\DATA\GESS\AffyCEL_Ex1_2.ges
3       345           ...\DATA\GESS\AffyCEL_Ex1_3.ges
4       345           ...\DATA\GESS\AffyCEL_Ex1_4.ges
5       345           ...\DATA\GESS\AffyCEL_Ex1_5.ges
6       345           ...\DATA\GESS\AffyCEL_Ex1_6.ges
```

This report shows a list of the output files. These are the names of the files that will be used as input for statistical analyses. The file also reports the total number of probe sets from each chip.

Row

This is the row of the array on the spreadsheet.

Number of Probe Sets

This contains the total number of probe sets processed for each chip.

File Name

This contains the path and name of each output (.ges) file.

Numerical Summaries

```
Numerical Summary of Original PM Intensities

                        10th        25th        50th        75th        90th
Row     Minimum   Percentile  Percentile  Percentile  Percentile  Percentile     Maximum
1          63.3        101.5       113.3       134.8       194.5       617.4     33464.8
2         114.3        151.8       163.3       184.8         245      711.15     24453.8
3          96.3        126.5         138       158.8       217.3       669.4     30724.3
4            87        125.8       137.8      159.15         224       676.8     38626.5
5            72        111.3       123.5       143.8       207.5      707.15     43899.8
6         137.3        176.3         188       209.4      266.35       713.4       27336

Numerical Summary of Expression Values (Log2 Scale)

                        10th        25th        50th        75th        90th
Row     Minimum   Percentile  Percentile  Percentile  Percentile  Percentile     Maximum
1      3.346565     3.793598    3.965115    4.271953    4.606102    5.031637    6.468214
2      3.320765     3.732144    3.972248    4.274407    4.593267    4.999708    7.798704
3      3.250465     3.752393    4.001728    4.232908    4.610593    5.058788    7.698844
4      3.183466     3.742249    3.959513    4.236988    4.668025    5.111084    6.274854
5      3.221636     3.673672    3.924484    4.264726    4.609824    5.034116    7.315955
6      3.372372      3.73401    3.965112    4.234389    4.588578    4.999599     6.44462
```

This report gives summary statistics of the original PM intensities and expression values for each chip.

Row

This is the row of the array in the spreadsheet.

Minimum

This is the minimum PM intensity or expression value.

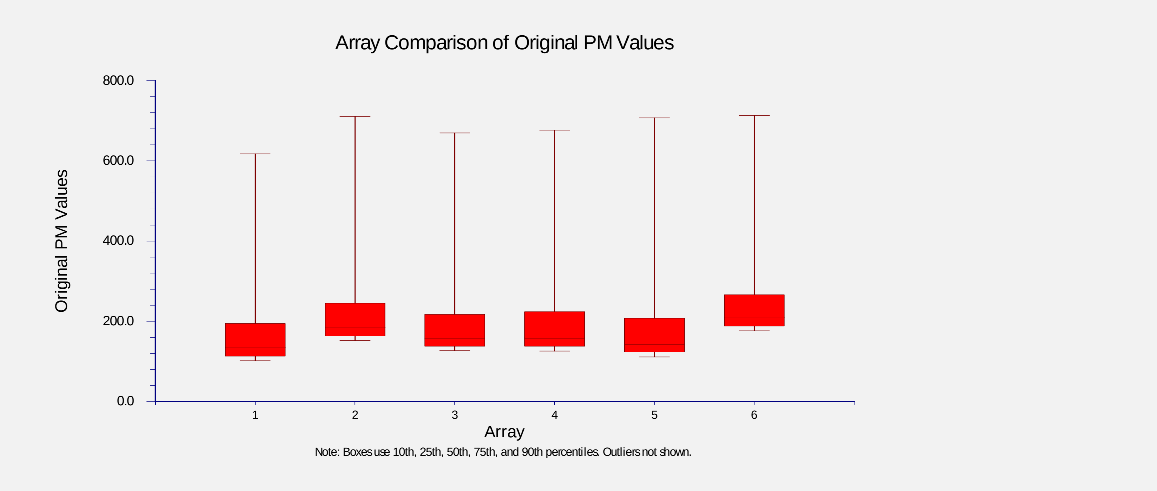
Percentiles

These are the 10th, 25th, 50th, 75th, and 90th percentiles of PM intensities or expression values.

Maximum

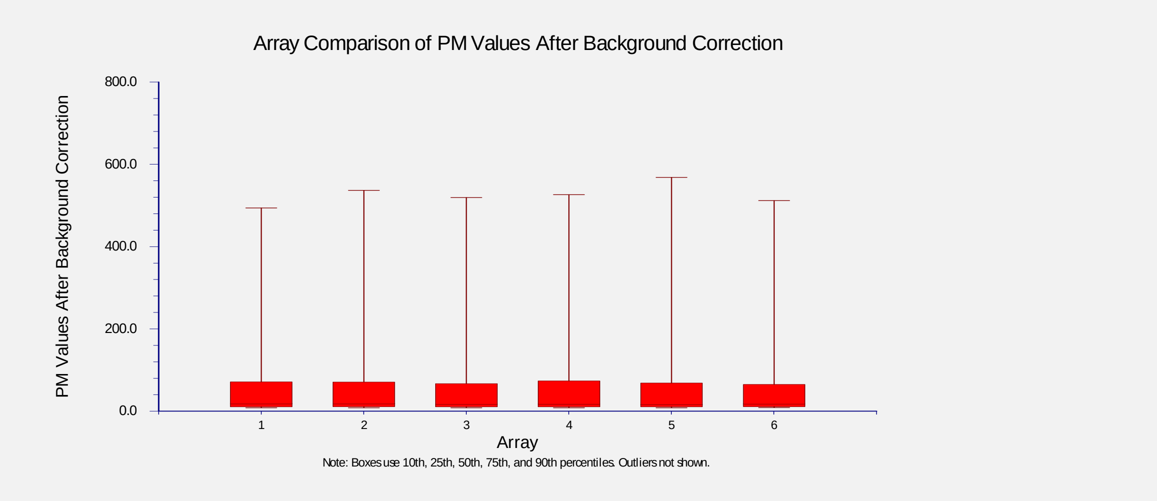
This is the maximum PM intensity or expression value.

Array Comparison of Original PM Values



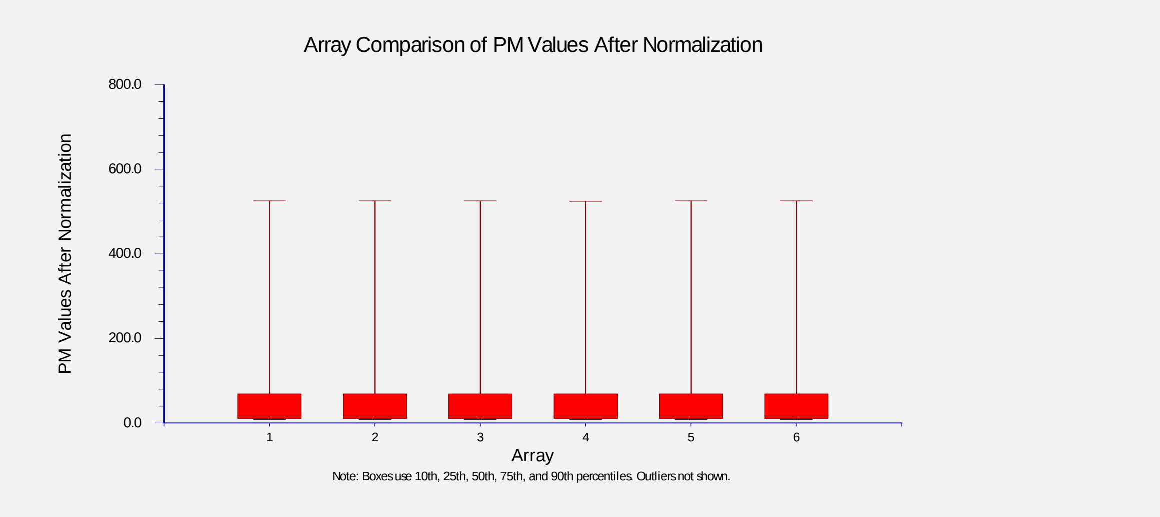
This plot shows the relative distributions of the original PM values.

Array Comparison of PM Values After Background Correction



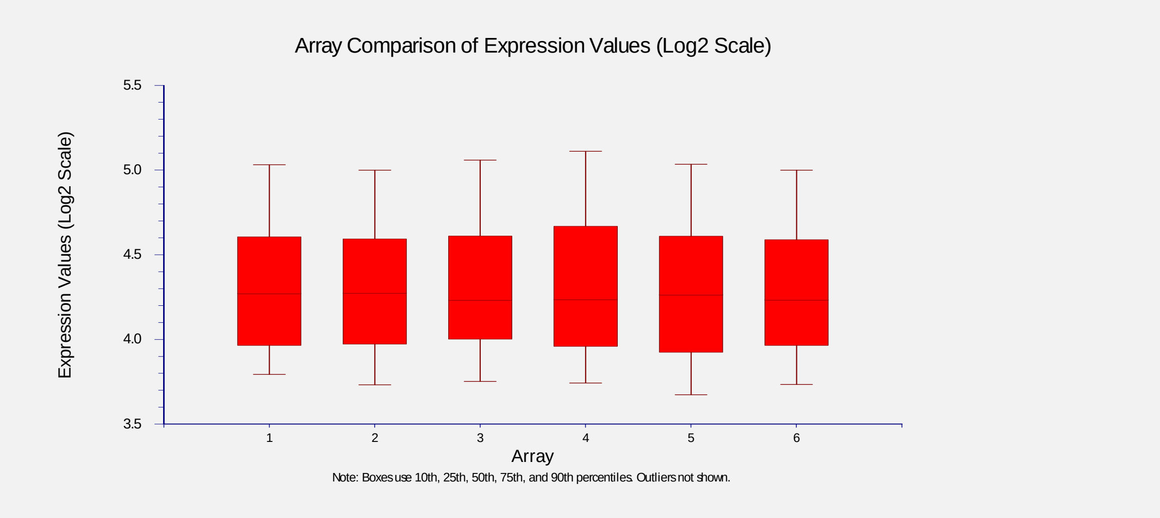
This plot shows the relative distributions of PM values after background correction. The values are shifted nearer to zero.

Array Comparison of PM Values After Normalization



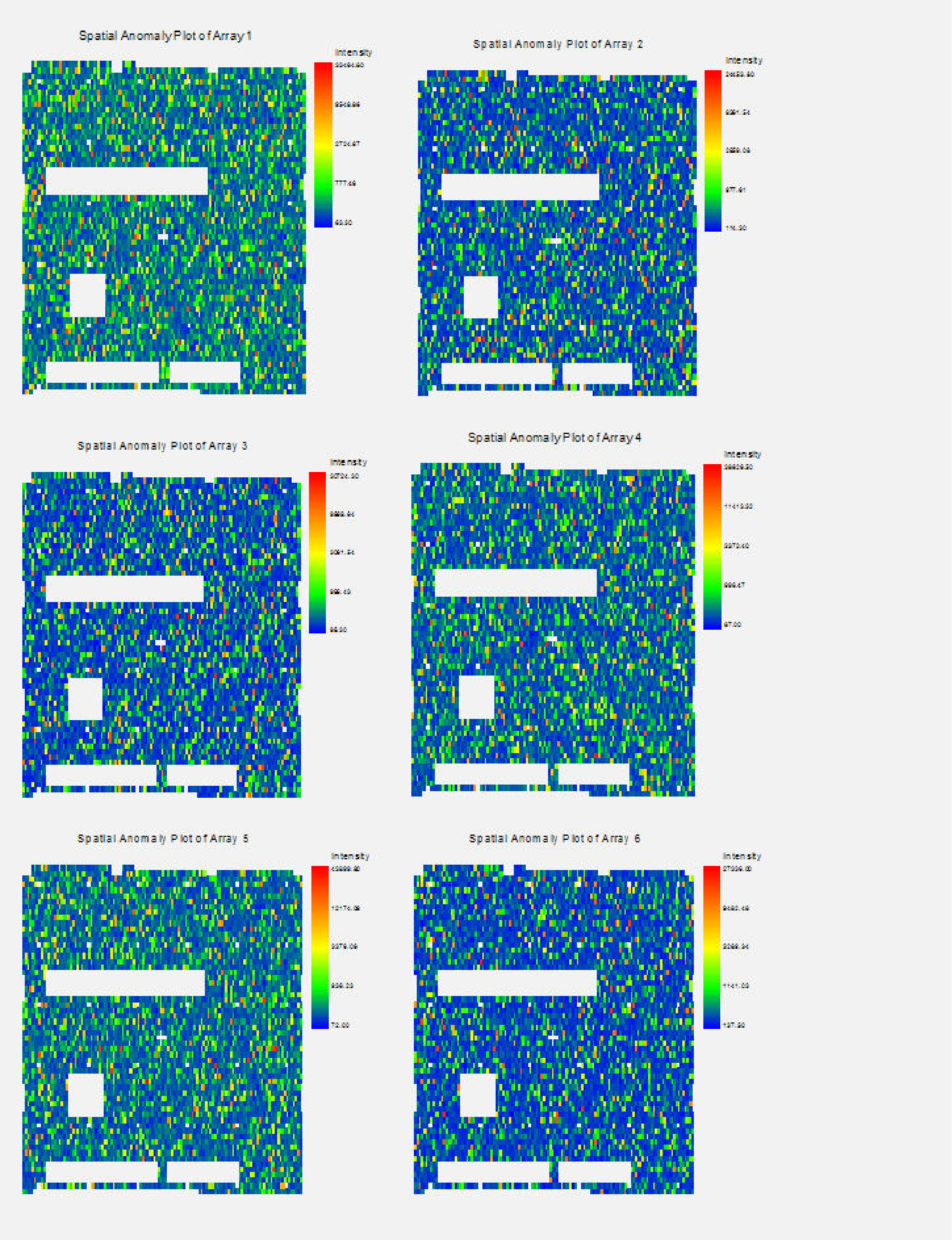
This plot shows the relative distributions of PM values after normalization. All chips now have the same distribution of PM values.

Array Comparison of Expression Values



This plot shows the relative distributions of expression values after summarization.

Spatial Anomaly Plots



This report shows a spatial representation of the original PM intensities. All six chips demonstrate random spatial distribution of intensities. The void areas correspond to Affymetrix QC probe cells that are not plotted.

Example 2 – Obtaining Quality Control Reports and Plots Only

This section presents an example of how to obtain numerical summaries, spatial anomaly plots, and a comparative box plot of original PM values without proceeding to background correction, normalization, and median polish. The spreadsheet data used are recorded in the AFFYCEL dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF* directory.

To run this example, take the following steps or load the **Example 2** template from the Affymetrix CEL File Pre-Processing Engine Template tab.

1 Open the AFFYCEL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS**.
- Click on the file **AFFYCEL.S0**.
- Click **Open**.

2 Open the Affymetrix CEL File Pre-Processing Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Affymetrix CEL Files**. The Affymetrix CEL File Pre-Processing Engine procedure will be displayed.
- On the Affymetrix CEL File Pre-Processing Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Affymetrix CEL File Pre-Processing Engine window, select the **Variables tab**.
- Set the **CDF File Name Variable** to **CDFFile**.
- Set the **CEL File Names Variable** to **InputFile**.
- Check the box next to **Generate Quality Control Reports and Plots Only**.
- Leave all other options under the Variables tab and other tabs at their default settings.

4 Specify the reports.

- Select the **Reports tab**.
- Check the box next to **Spatial Anomaly Plots**.
- Leave all other options under the Reports tab at their default settings.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Only the **Input Files Summary**, **Numerical Summary of PM Intensities**, **Array Comparison of Original PM Values** box plot, and **Spatial Anomaly Plots** will be displayed. No pre-processing will be performed and no output files are written. This allows you to inspect chips before proceeding.

Chapter 111

Affymetrix CHP File Import Engine

Unlike .cel files, which contain intensity measures, Affymetrix .chp files contain expression values (signal estimates) calculated by Affymetrix microarray processing software. This chapter describes the process of importing expression values directly from .chp files using the ***GESS*** Affymetrix CHP File Import Engine. Before performing statistical analysis on data stored in .chp files, you must convert each .chp file into a .ges file, the type of file read by ***GESS*** analysis routines. Compatible .chp expression files are those generated by Affymetrix MAS 5.0 and above or Affymetrix GCOS 1.X software.

The engine works by taking each .chp file from the designated column on the spreadsheet and creating an output (.ges) expression file. As is the case with the processing of Affymetrix .cel files, a .cdf library file is required to interpret the .chp file and must be entered in the first cell of an empty column on the spreadsheet. Multiple .chp files may be imported on a single run of the ***GESS*** Affymetrix CHP File Import Engine. For more information about Affymetrix microarray technology, see the Affymetrix CEL File Pre-Processing Engine chapter in this manual or visit www.affymetrix.com.

Entering CDF and CHP Files

This section describes how file names are entered into the spreadsheet in preparation for importing .chp files. Three variables (columns) are required for the import process: two designate the input files, and the third receives the .ges file names for further analysis. Any name (entered by clicking on the Variable Info tab at the bottom of the spreadsheet) may be used for these variables.

CDF File Name Variable

The CDF File Name Variable is a column on the spreadsheet containing the filename and path of the .cdf file that corresponds to the chips from which the .chp files were created. The name and path must be entered in the **first cell** of this column. This variable is required to run the Affymetrix CHP File Import Engine.

To enter the .cdf file name and path into the spreadsheet:

1. Highlight the first cell in an empty column.
2. Either type in the path and file name directly or hit **F7** to browse for the appropriate .cdf file.

CHP File Names Variable

The CHP File Names Variable is a column on the spreadsheet containing a list of filenames and paths of the .chp files that are to be imported. The files may be in different folders but all must correspond to the same array type (.cdf file). This variable is required to run the Affymetrix CHP File Import Engine.

To enter a .chp file name and path into the spreadsheet:

1. Highlight an empty cell.
2. Either type in the path and file name directly or hit **F7** to browse for an appropriate .chp file.
3. Repeat steps 1 and 2 until all .chp files have been entered.

Output File Names Variable

When the import engine is run, a new set of files is generated automatically for use in statistical analyses. These output files have the extension .ges, and are stored in the folder specified under Folder in which Output Files will be Stored. The path and name of each .ges output file is placed on the spreadsheet in the column specified by the Output File Names Variable. This column of output files will become your input files for further statistical analyses. This variable is required to obtain output for statistical analysis.

To specify an output folder under Folder in which Output Files will be Stored:

1. Click on the **Browse** button to the right of the window.
2. Select an output folder and click **OK**.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options specify the variables and major file-importing options that will be used in this procedure.

Affymetrix Files for Importing

These options are used specify the .cdf and .chp files that are to be used during the import process.

CDF File Name Variable

Select the variable that contains the .cdf file corresponding to the .chp files to be imported.

<u>The name and path of the file must appear in the **first cell** of the column below this variable name on the spreadsheet.</u> To enter the .cdf file name and path into the spreadsheet:

1. Highlight the first cell in an empty column.
2. Either type in the path and file name directly or hit F7 to browse for the appropriate .cdf file.

When a .cdf file is first used in pre-processing, ***GESS*** will automatically create and store a new file containing the .cdf file information. This new file is stored in the *C:\…\[My] Documents \NCSS\NCSS 2007\Data\GESS\CDH* folder and has the extension ".cdh". ***GESS*** reads the .cdh file much faster than it reads the corresponding .cdf file. In subsequent runs of the Affymetrix Pre-Processing Engine, ***GESS*** will preferentially use the stored .cdh file to increase efficiency.

A new .cdh file is stored for each different .cdf file used. For example, if you processed .cel files using HG-U133A.cdf, a new file with the name "HG-U133A.cdh" would be created. On subsequent runs using HG-133A chips, the HG-U133A.cdh file will be used. If at some point you processed new .cel files using HG-Focus.cdf, the file "HG-Focus.cdh" would be stored. If you believe that a .cdf file you are using has changed since you last ran the engine, you should delete the current .cdh file from the *C:\…\[My] Documents\NCSS\NCSS 2007\Data\GESS\CDH* folder before running the procedure.

CHP File Names Variable

Select the variable that contains the list of the .chp files to be imported.

The names and pathways of the files should appear in a column below this variable name on the spreadsheet. To enter the .chp file names and paths into the spreadsheet:

1. Highlight an empty cell.
2. Either type in the path and file name directly or hit **F7** to browse for an appropriate .chp file.
3. Repeat steps 1 and 2 until all .chp files have been entered.

GES Output File Specifications

These options are used to specify the location and naming of the output .ges files.

Folder in which Output Files will be Stored

Enter the path and name of the folder in which the newly created .ges files will be stored. The path may be typed directly, or the browse button may be used to locate the desired folder.

In the default path "%mydocs_NCSS%\DATA\GESS", the designator "%mydocs_NCSS%" represents the path to the ***GESS*** personal data folder (commonly *C:\...\[My] Documents\NCSS \NCSS 2007*).

Output File Names Variable

Select the variable in which the path and names of the newly created .ges files will be stored for future statistical analyses. The path and folder for these .ges files is entered under Folder in which Output Files will be Stored.

A new file (to be used for statistical analyses) will be created for each input array when the procedure is run.

CAUTION: If a variable containing data is entered, the data on the spreadsheet will be overwritten and lost.

Append to File Names

A sequence of characters may be entered here to append to the name of the created file. For example, if the .chp file has the name "Slide1_10hours.chp" and "log" is entered here, the newly created .ges file will be "Slide1_10hours log.ges".

If nothing is entered here, the file name will be the same as the name of the .chp file, but ".chp" will be replaced with ".ges".

Overwrite existing output (.ges) files with new output (.ges) files

Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file. For example, if "Slide1.ges" has already been created and a new "Slide1.ges" file is to be written, the new file will be "Slide1 (2).ges" if the Overwrite box is not checked.

Transformation Options

These options determine the expression value that will be stored in the output files.

Data Transformation

Specify the transformation to perform on the input expression data before saving output files. This option allows you to replace negative numbers and/or perform logarithmic transformations on the data. If a log transformation is chosen, negative numbers and zeros in the data must be replaced by either missing values or the Replacement Value before the log transformation is computed. The available options are:

- **None**

  No transformation is performed on the data.

- **Set Negative Numbers to Zero**

  Negative numbers are replaced with zeros. No log transformation is performed.

- **Set Negative Numbers to the Replacement Value**

  Negative numbers are replaced with the Replacement Value. No log transformation is performed.

- **Set Negative Numbers to the Missing Values**

  Negative numbers data are replaced with missing values. No log transformation is performed.

- **Log Base X (Set Negative Numbers and Zeros to the Replacement Value)**

  Negative numbers and zeros are replaced with the Replacement Value prior to taking the log of the data. The log options are log base 2, log base e (natural log), and log base 10.

- **Log Base X (Set Negative Numbers and Zeros to the Missing Values)**

  Negative numbers and zeros are replaced with missing values prior to taking the log of the data. The log options are log base 2, log base e (natural log), and log base 10.

Replacement Value

Specify the value that will replace negative values and/or zeros in the dataset. This option is only used when the "Data Transformation" selected requires the use of the Replacement Values. Otherwise, this option is ignored.

RANGE: Replacement Value > 0

Reports Tab

The options on this panel control which reports and plots are generated.

Select Reports

The following reports and report options are available.

File Processing Summary

Check this box to obtain a row-by-row summary of input and output file names and a summary of input and output file specifications.

Data Summary

Check this box to obtain a numeric summary of the data saved in the newly created .ges files.

Decimals

Specify the number of decimals to be used for percentiles in the Data Summary report. The number of decimal places is used for output display only. It does not change the internal precision of the data.

Select Plots

Choose from the following plots.

Comparative Box Plot

Check this box to obtain a comparative box plot of expression values.

Box Plot Tab

The options on this panel control attributes of the box plot.

Horizontal and Vertical Axes

The following options allow you to format the horizontal (X) and vertical (Y) axes.

Label

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate X-axis label.

{Y} is replaced by an appropriate Y-axis label.

Ref. Number Format...

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Minimum

Specify the value to be displayed as the minimum on this axis. If this value is left blank, the minimum will be determined from the data. If this value is greater than the smallest 10th percentile of the data, it will be ignored.

Maximum

Specify the value to be displayed as the maximum on this axis. If this value is left blank, the maximum will be determined from the data. If this value is less than the largest 90th percentile of the data, it will be ignored.

Major Ticks

Specify the number of large tickmarks and optional grid lines along the axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along the axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along the axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the corresponding axis. Thus, checking Show Grid Lines here will actually cause horizontal grid lines to appear.

Box Plot Settings

The following options allow you to control the box plot.

Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created in the Box Plots procedure.

% Space

When the Box Width (or Bar Width) option is set to Percent Space in the box plot style file selected, this value specifies the percent of the length of the axis that is empty space instead of bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also, note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**
  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**
  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Interior

The color used to fill the rectangle formed by the vertical and horizontal axes. Click to change.

Background

The color used behind the plot. Click to change.

Box Fill

The color used to fill the boxes. Click to change.

Box Border

The color used to outline the boxes. Click to change.

Line

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Top and Bottom Titles

Enter text here for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate horizontal-axis label.

{Y} is replaced by an appropriate vertical-axis label.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Importing Affymetrix CHP Files

This section presents an example of how to import two .chp expression files. The spreadsheet data used are recorded in the AFFYCHP dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\AF* directory.

To run this example, take the following steps or load the **Example 1** template from the Affymetrix CHP File Import Engine Template tab.

1 Open the AFFYCHP dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **AFFYCHP.S0**.
- Click **Open**.

2 Open the Affymetrix CHP File Import Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Affymetrix CHP Files**. The Affymetrix CHP File Import Engine procedure window will be displayed.
- On the Affymetrix CHP File Import Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

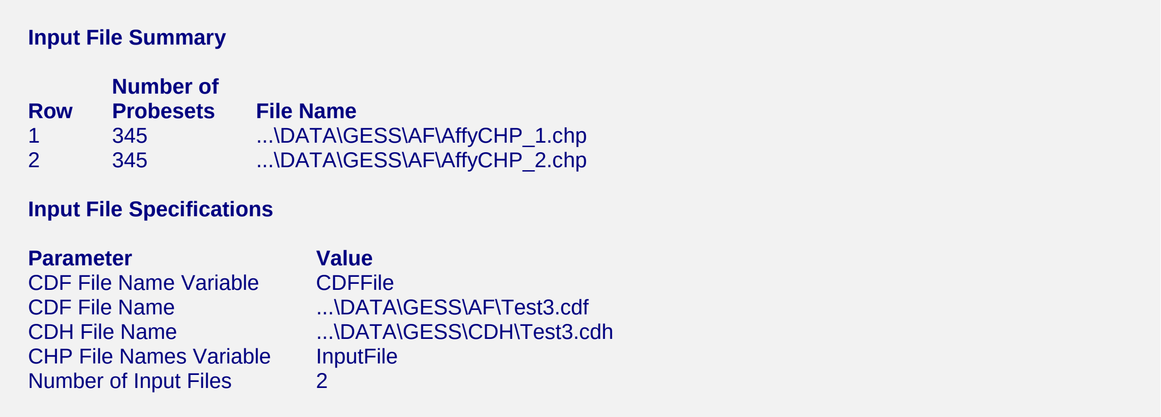
3 Specify the variables.

- On the Affymetrix CHP File Import Engine window, select the **Variables tab**.
- Set the **CDF File Name Variable** to **CDFFile**.
- Set the **CEL File Names Variable** to **InputFile**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **OutputFile**.
- Put a check next to **Overwrite existing output (.ges) with new output (.ges) files**.
- Leave all other options under the Variables tab and other tabs at their default settings.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Input File Summary



```
Input File Summary

        Number of
Row     Probesets     File Name
1       345           ...\DATA\GESS\AF\AffyCHP_1.chp
2       345           ...\DATA\GESS\AF\AffyCHP_2.chp

Input File Specifications

Parameter                    Value
CDF File Name Variable       CDFFile
CDF File Name                ...\DATA\GESS\AF\Test3.cdf
CDH File Name                ...\DATA\GESS\CDH\Test3.cdh
CHP File Names Variable      InputFile
Number of Input Files        2
```

This report displays a list of input files and the number of probesets contained in each input file. This report also lists the input file specifications used.

Row

This is the row of the input file on the spreadsheet.

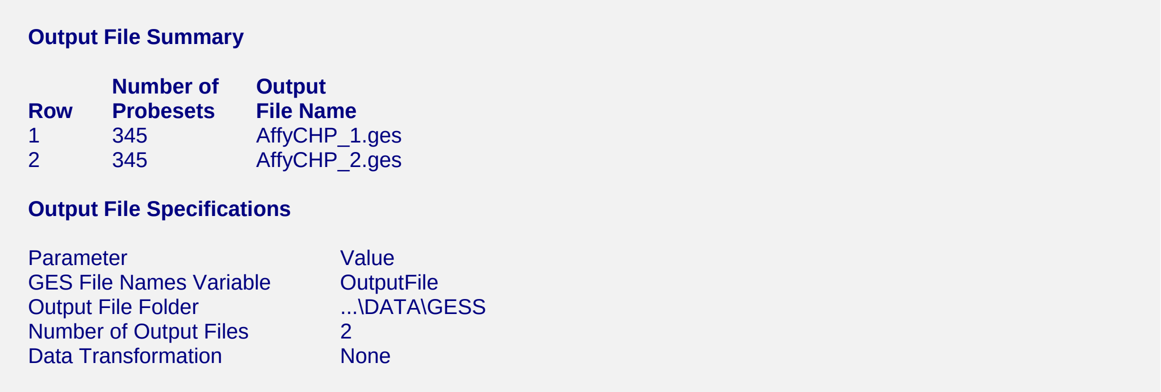
Number of Probesets

This is the number of probesets stored in each input file.

File Name

This contains the path and name of each input file.

Output File Summary



```
Output File Summary

        Number of     Output
Row     Probesets     File Name
1       345           AffyCHP_1.ges
2       345           AffyCHP_2.ges

Output File Specifications

Parameter                      Value
GES File Names Variable        OutputFile
Output File Folder             ...\DATA\GESS
Number of Output Files         2
Data Transformation            None
```

This report displays a list of the output file names and the number of genes contained in each output file. A list of output file specifications is also given.

Row

This is the row of the newly created .ges file on the spreadsheet.

Number of Probesets

This is the number of probesets stored in each input file.

Output File Name

These are the names of the newly created .ges files. The folder into which these files were stored is listed as the "Output File Folder".

Numerical Summary of Expression Values

```
Numerical Summary of Expression Values

                     10th       25th       50th       75th       90th
Row  Minimum   Percentile Percentile Percentile Percentile Percentile  Maximum
1    2.799185    10.10108   22.67343   72.04096   243.2257   1634.838  63009.21
2    1.998915     8.33041   18.69839   60.80057   236.5944    1470.74   57153.7
```

This report gives numerical summaries of the expression values saved in the output files.

Row

This is the row of the output file on the spreadsheet.

Minimum

This is the minimum expression value stored.

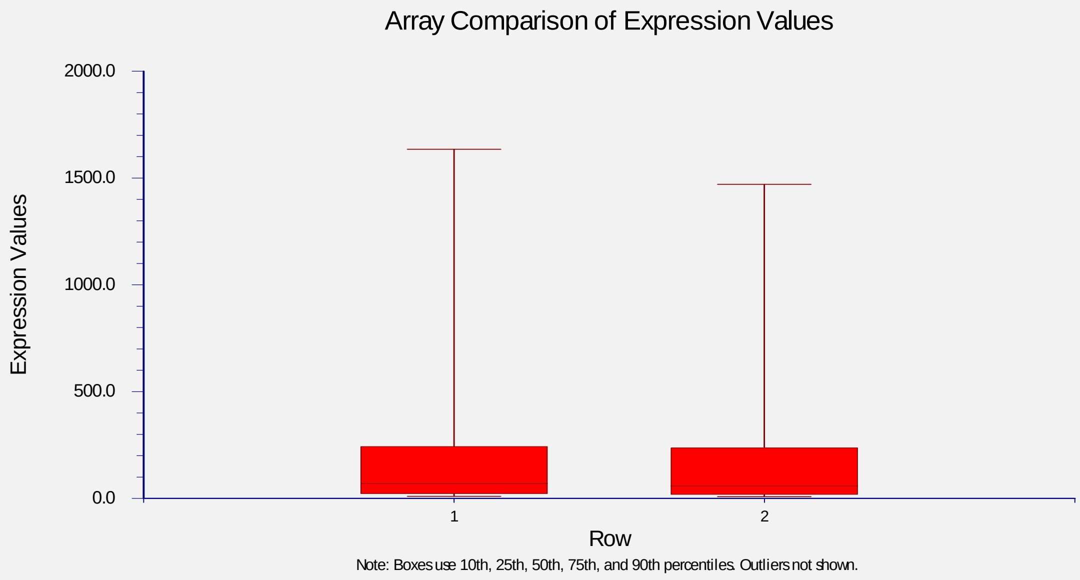
Percentiles

These are the 10th, 25th, 50th, 75th, and 90th expression value percentiles.

Maximum

This is the maximum expression value stored.

Array Comparison of Expression Values



This plot shows the relative distributions of expression values.

Example 2 – Performing a Data Transformation

This section presents an example of how to import two .chp expression files and perform a logarithmic transformation on the data before saving the .ges output file. The spreadsheet data used are recorded in the AFFYCHP dataset. The input files are stored in the *C:\Program Files \NCSS\NCSS 2007\Data\GESS\AF* directory.

To run this example, take the following steps or load the **Example 2** template from the Affymetrix CHP File Import Engine Template tab.

1 Open the AFFYCHP dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your NCSS directory.
- Open the **GESS** folder.
- Click on the file **AFFYCHP.S0**.
- Click **Open**.

2 Open the Affymetrix CHP File Import Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Affymetrix CHP Files**. The Affymetrix CHP File Import Engine procedure will be displayed.
- On the Affymetrix CHP File Import Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the Variables.

- On the Affymetrix CHP File Import Engine window, select the **Variables tab**.
- Set the **CDF File Name Variable** to **CDFFile**.
- Set the **CEL File Names Variable** to **InputFile**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **OutputFile**.
- Under **Append to File Names**, enter **log2**.
- Put a check next to **Overwrite existing output (.ges) with new output (.ges) files**.
- Under **Data Transformation**, select **Log base 2 (Set Negative Numbers and Zeros to the Replacement Value)**.
- Leave all other options under the Variables tab and other tabs at their default settings.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Input File Summary

The input file summary is the same as that in Example 1.

Output File Summary

```
Output File Summary

       Number of     Output
Row    Probesets     File Name
1      345           AffyCHP_1 log2.ges
2      345           AffyCHP_2 log2.ges
```

The extension "log2" has been added to the end of the .ges file names.

Transformation Summary

```
Transformation Summary
Data Transformation: Log base 2 (Set Negative Numbers and Zeros to the Replacement Value = 0.001)

       Number of    Number of
       Negatives        Zeros     Output
Row     Replaced     Replaced     File Name
1              0            0     AffyCHP_1 log2.ges
2              0            0     AffyCHP_2 log2.ges
```

This report is only obtained if a data transformation of some type is performed. The report indicates that neither zeros nor negative values were encountered.

Row

This is the row of the output file on the spreadsheet.

Number of Negatives Replaced

This is the total number of negative values that were encountered and replaced for each .chp file.

Number of Zeros Replaced

This is the total number of zeros that were encountered and replaced for each .chp file.

Output File Name

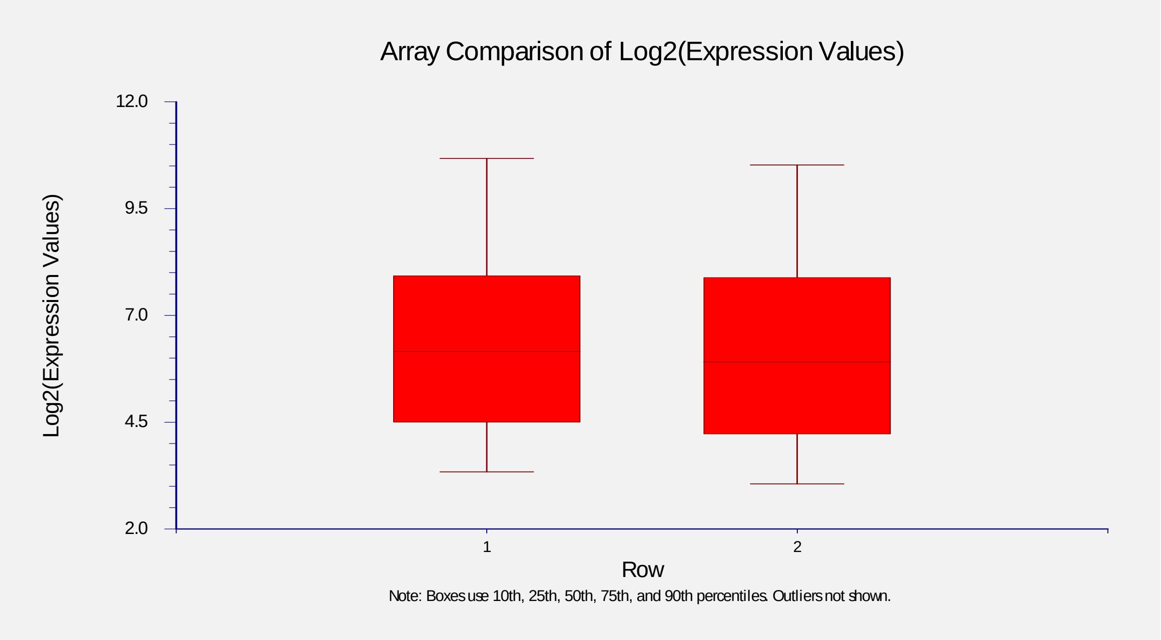
These are the names of the newly created .ges files. The folder into which these files were stored is listed as the "Output File Folder" in the Output File Summary.

Numerical Summary of Expression Values

```
Numerical Summary of Log2(Expression Values)

                         10th         25th         50th         75th         90th
Row    Minimum     Percentile   Percentile   Percentile   Percentile   Percentile     Maximum
1     1.485007       3.336433      4.50293     6.170745     7.926146      10.6749    15.94328
2    0.9992173        3.05835     4.224842     5.926013     7.886229     10.52227    15.80256
```

These are the summary statistics of the expression values that were stored for each .chp file. These values are on the log base 2 scale.

Array Comparison of Expression Values



This plot shows the relative distributions of expression values on the log base 2 scale.

111-14 Affymetrix CHP File Import Engine

Chapter 120

Agilent® TXT File Pre-Processing Engine

Introduction

The main purpose of this chapter is to describe the process of obtaining relative expression values from Agilent® Feature Extraction output using the ***GESS*** Agilent TXT File Pre-Processing Engine. The Agilent® Feature Extraction software is used to summarize the pixel information of a high resolution image. Each column of the resulting (.txt) file provides a name, description, or numeric attribute for each feature (spot) of the array.

Following a brief background to the concept of microarrays, this chapter discusses many of the principles and practical aspects of two-channel arrays, including array production, image analysis, array and spot quality, filtering issues, and normalization. Reference and paired experimental designs are also presented as well as the dye-swap technique. The chapter concludes with a tutorial of the entire process of using the Agilent TXT File Pre-Processing Engine to obtain expression values from Agilent (.txt) file output.

Chapter Structure

Background

An overview of microarray concepts is presented first. This section is designed to familiarize a non-biologist with the concepts of DNA expression and microarray hybridization.

Nine Steps to Obtain Relative Expression Values

The background is followed by a summary of the nine steps required to obtain final relative expression values for a single ***two-channel*** array, which can then be used in comparison analysis.

Step 1 – *In Situ* Microarray Fabrication. Probe sequences are generated for each probe using inkjet technology.

Step 2 – Hybridization. Two samples are each labeled with a different dye, mixed, and then introduced onto the array. The sequences that are complementary to each probe will bind to the probe sequences. The two samples compete for binding at each probe.

Step 3 – Array Image Construction. The array is scanned twice, once for each dye, producing two image files. The image files contain a value (representing a shade of gray) for every pixel on the array.

Step 4 – Array Image Processing. Image processing software is used to locate each spot and separate foreground and background regions for both dyes.

Step 5 – Image Quantification. Image processing software produces summaries of the foreground and background pixels for both dyes at each spot. Relative intensities, comparing red (Cyanine 5) to green (Cyanine 3) dyes, are also produced.

Step 6 – Array Spot Types. Results from specially designed probes can be used to assess array quality.

Step 7 – Whole Array Quality. Specialized plots and numeric summaries of the whole array give indication of possible dye bias, spatial variation, or other artifacts that indicate whether values obtained from the array can be trusted for comparison.

Step 8 – Array Individual Spot Quality and Filtering. Pixel summaries and other values for each individual spot give indication of the spot quality. A variety of filters can be used to remove spots of questionable quality.

Step 9 – Array Normalization. A whole array normalization is recommended to correct for dye bias.

The result of these nine steps is a column of relative expression values that can be compared to corresponding values of other arrays in the experiment.

Agilent® Feature Extraction (.txt) Files

The results (.txt) file, obtained using Agilent® Feature Extraction image analysis software, is described.

Two-Channel Designs

Paired and Reference Experimental Designs are described, as well as the dye-swap technique.

Entering Agilent (.txt) Files

Details of entering Agilent (.txt) files into the spreadsheet are explained.

Procedure Options

The options available in ***GESS*** for preprocessing Agilent (.txt) files are described in detail.

Tutorial/Examples

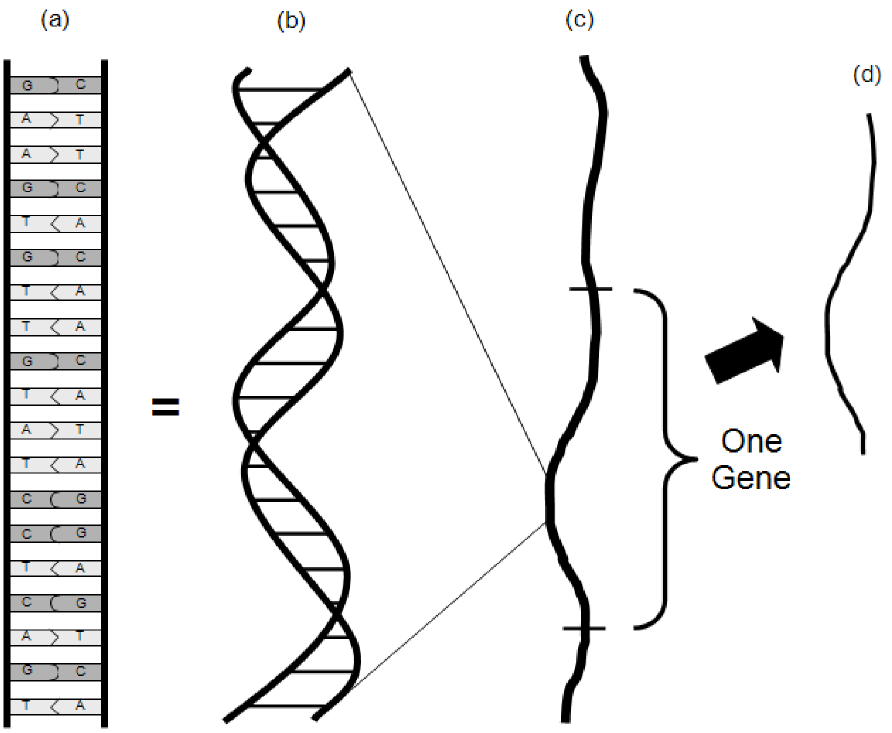
Examples of pre-processing Agilent (.txt) files in ***GESS*** are shown.

Background

Gene Expression

The general process of gene expression in a cell begins with the DNA. Each DNA molecule has the well-known double helix design. Each rung or step of a DNA molecule is made up of two *nucleotides*. The two nucleotides bonded together are called *base pairs*. In DNA the 4 possible nucleotides are A, T, C, and G (for Adenine, Thymine, Cytosine, and Guanine, respectively). The nucleotide A can only form a base pair with T, and vice versa. Similarly, C can only bind to G, and vice versa. A gene is a unique segment of a DNA molecule consisting of a series of base pairs ranging from about 50 to thousands of base pairs in length. When the need for a specific protein in the cell is identified, the gene for that protein is "read" and a *messenger RNA* (mRNA) is produced in a process called *transcription*. mRNA molecules are single-stranded molecules which are essentially copies of the gene segment of one of the two DNA strands. The mRNA is then used to produce a protein that is specific to that mRNA molecule in a process called translation.

DNA overview. (a) A ladder representation of a DNA segment showing 20 complementary base pairs. (b) A drawing of the three-dimensional form of the corresponding double helix. (c) The 20 base pairs of (a) and (b) are only a small section of the total DNA double strand. A gene is a segment of the DNA helix that contains the code for the production of a protein. (d) A single stranded mRNA molecule is generated when the gene is expressed. The mRNA molecule will be used to produce a protein that is specific to the gene of (c).

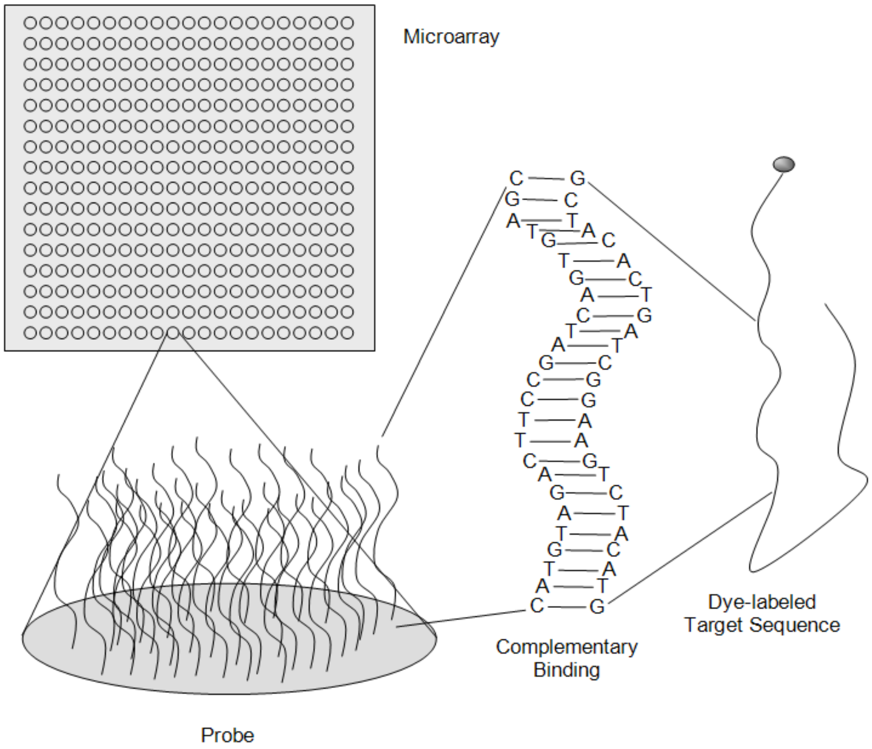


The newly created protein can then be used in the cell to perform the needed function. A gene that is in the process of producing or has produced a protein is said to be *expressed*. Expression of a given gene at a given time can thus be measured either by the amount of mRNA or protein (corresponding to that gene) in the cell. Microarrays are currently the prominent tool for quantifying the amount of mRNA in the cell (or collection of cells) for hundreds or thousands of genes simultaneously.

The Microarray

On a typical microarray, there are several thousand spots with *probes* (see below) of known identity, with each probe corresponding to a gene of interest. The probe sequences on each spot are designed to attract only sequences that are expressed by the gene to which that spot corresponds. A spot with the attached probe sequences may collectively be called a probe.

Below is a microarray drawing depicting the arrangement of spots on a spotted array (top left), the identical probe sequences on an individual spot (bottom left), and a segment of the probe for this spot that uniquely attracts the mRNA (or cDNA, a more stable mRNA replicate) strands of interest (center and right).



Hybridization (Binding to the Microarray)

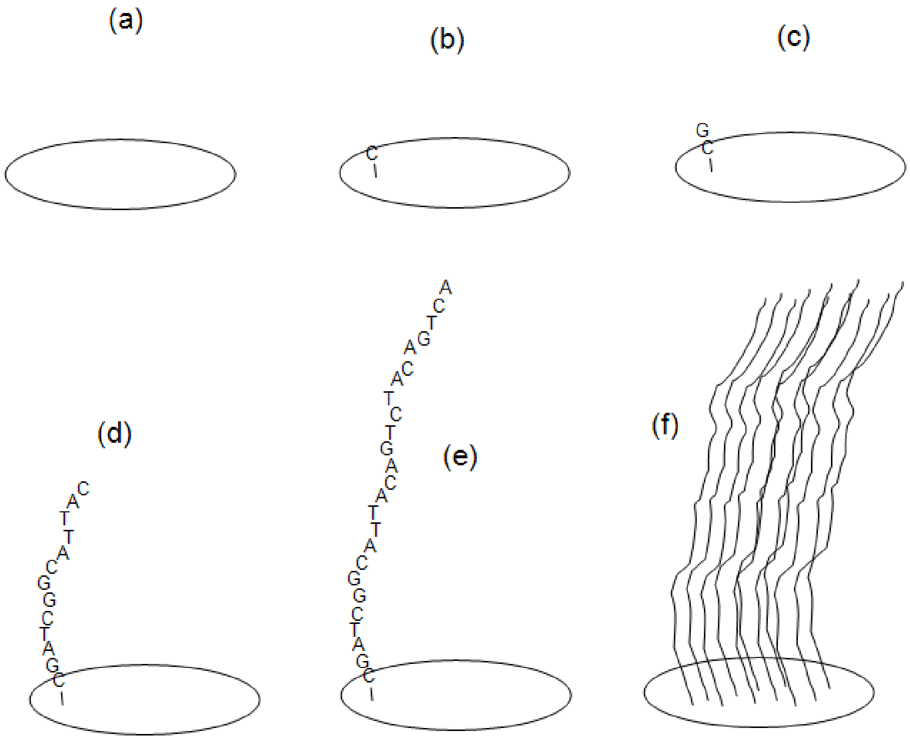
The mRNA expressed in an experimental unit is obtained from some of the experimental unit’s cells (i.e., blood or tissue), converted to cDNA (a nearly equivalent, but more stable molecule) using a process called reverse transcription, and labeled with fluorescent dye. When the solution containing the cDNA is exposed to a microarray, each of the cDNA sequences will bind to the probe sequences to which it complements. Thus, only sequences with perfect complementation along the entire sequence should bind to the corresponding probe (see figure above). The cDNA from genes which are expressed in higher quantities will hybridize (bind) to the corresponding probe in higher quantities. The amount of hybridized material for each spot can then be measured using the intensity of fluorescence from the bound cDNA when exposed to laser scanning. A scanner (or scanning machine) measures the intensity of fluorescence for every spot on the array. The result of a single microarray scan is several thousand intensities representing the amount of mRNA expression of those genes that are probed on the array.

Nine Steps to Obtain Relative Expression Values

Step 1 – In Situ Microarray Fabrication

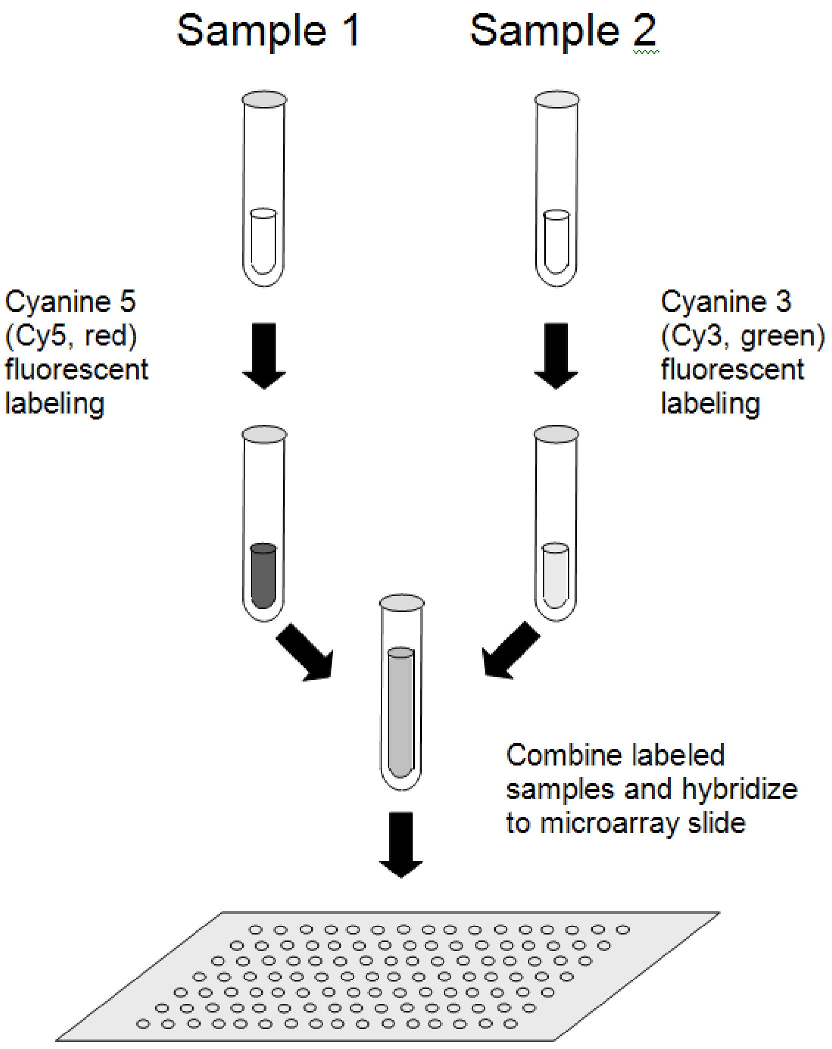
With *in situ* microarray synthesis, probe sequences are constructed nucleotide by nucleotide, directly on the array. A general advantage of this method of probe synthesis is that the sequence of every probe is known exactly. A disadvantage is that *in situ* synthesis techniques limit the probe sequences to lengths much shorter than those of spotted probes.

Agilent® oligo microarrays are produced using an inkjet printing process to attach individual nucleotides to the array. The length of the probe sequences used in Agilent® arrays is 60 nucleotides. Four cartridges, each containing one of the four nucleotides, are used to deposit the nucleotides in the correct location, based on digital sequence files. The progress of generating probe sequences is seen in the figure. Multiple copies of the same probe sequence are generated in a tiny circular region to form a single probe spot. Thousands of probes (spots) are synthesized onto a single array.


(a)
(b)
C
(c)
G
C
(d)
C G A T C G G C A T T A C
(e)
C G A T C G G C A T T A C A G T C T A C A G T C A
(f)

Step 2 – Hybridization

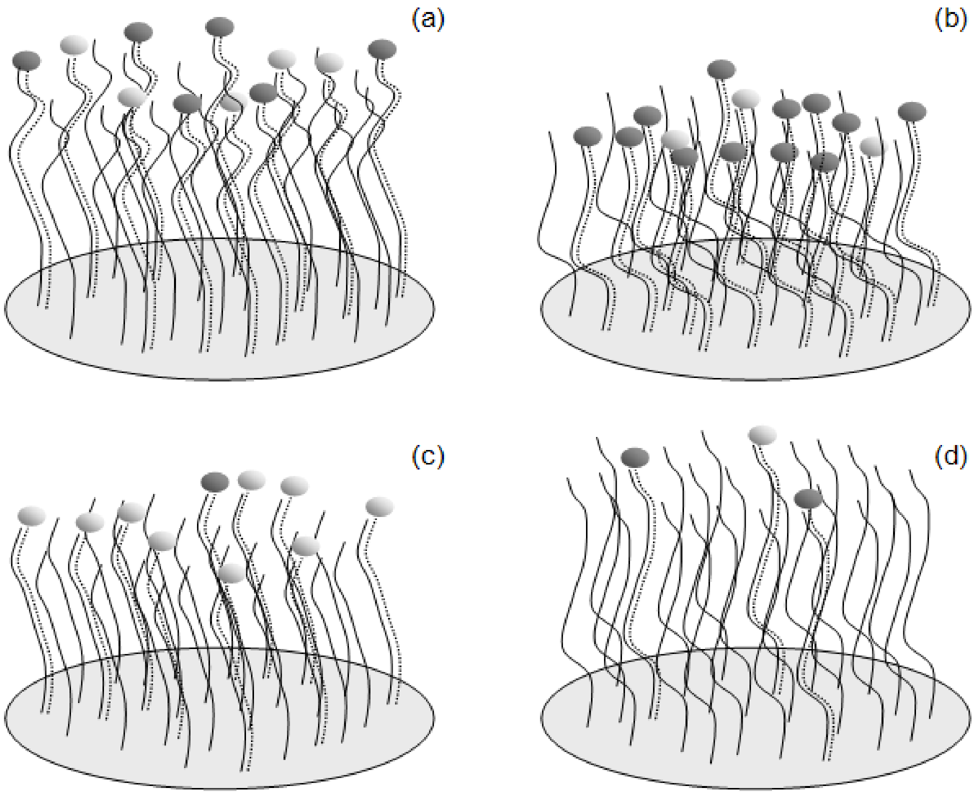
Two-channel microarrays refer to those for which two samples (and two dyes) are analyzed on each array. Two-channel microarrays are also called two-color microarrays. One cDNA sample is labeled with Cyanine 5 (Cy5, red) dye, and the other sample is labeled with Cyanine 3 (Cy3, green) dye. The samples are mixed and then introduced onto the array to compete for hybridization at each spot, as shown in the following diagram of the two-channel fluorescent labeling process.


Sample 1
Sample 2
Cyanine 5 (Cy5, red) fluorescent labeling
Cyanine 3 (Cy3, green) fluorescent labeling
Combine labeled samples and hybridize to microarray slide

If a specific sequence is highly expressed in one of the samples (say, Sample 1) and has low expression in the other sample (say, Sample 2), the probe for that sequence should bind more Sample 1 sequences than Sample 2 sequences. This process is called competitive hybridization.

Competitive Hybridization

Below are some examples of competitive hybridization at individual spots. The dotted line sequences with dark dots attached represent Sample 1 cDNA with Cyanine 5 (Cy5, red) fluorescent labels. The dotted line sequences with light dots attached represent Sample 2 cDNA with Cyanine 3 (Cy3, green) fluorescent labels. Each of the four examples show varying amounts of competitive hybridization. For the probe in (a), there is nearly equal expression among both channels. In (b), there is high expression of this gene in Sample 1, low expression in Sample 2. In (c) can be seen very low expression of this gene in Sample 1, but high expression in Sample 2. There is very low expression for this gene in both samples in (d).

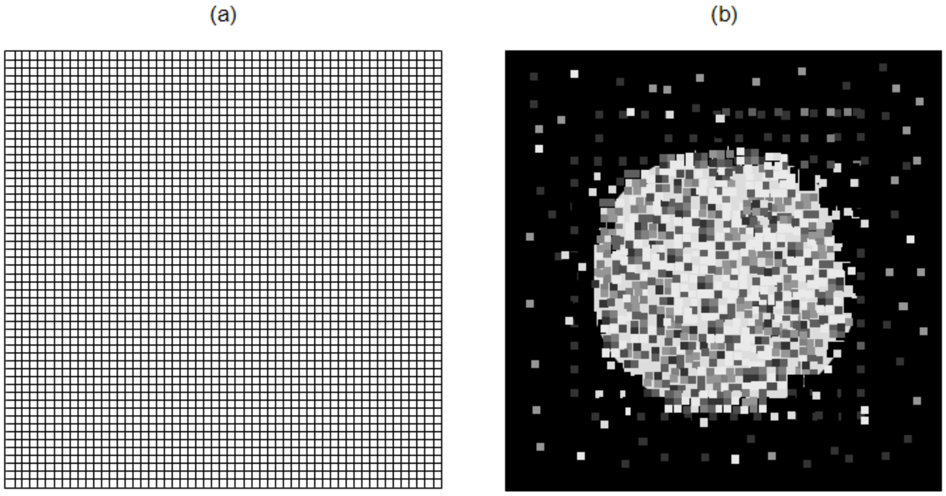


Step 3 – Array Image Construction

Following competitive hybridization, the laser of a scanning machine is used to illuminate the fluorescent dye of one of the channels (e.g., Cyanine 5) across the whole array, creating a high resolution black-and-white image for that channel. The frequency of the laser is then adjusted (or a different laser is used) to illuminate the fluorescent dye of the other channel (e.g., Cyanine 3), creating a second black-and-white image. Each of the images is usually stored as Tag Image File Format (.tif) file. The image is made up of a grid of pixels. Because each pixel is stored using 16 bits of memory, each pixel can take on any of 2^{16} = 65,536 shades of gray. The numeric range for each pixel is thus 0 to 65,535. The number of pixels in each spot depends upon the resolution (total number of pixels) of the image and the size of the spot (see the figure below).

Pixel Grid and .tif Image following Laser Scanning

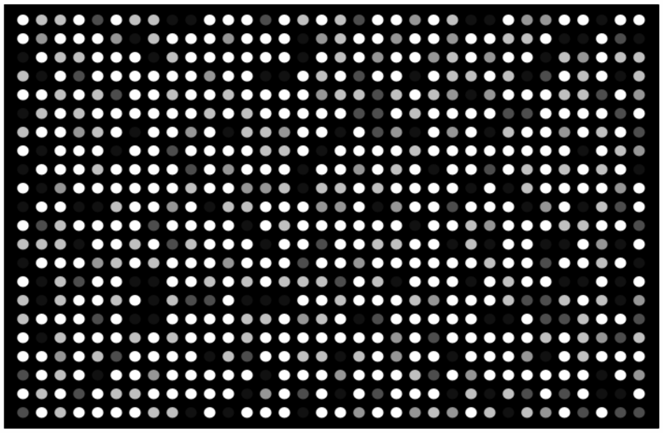
A pixel grid for the region of a single probe spot is shown in (a). A drawing of a .tif image following laser scanning is shown in (b). The number of pixels in a spot may range from below 50 to several hundred, depending upon the size of the spot and the resolution of the image.



Artificial colors (i.e., red and green) are often used in image construction software to visualize the expression represented on each array. The colors may be superimposed to form a single array image that displays the relative illumination at each spot. On the superimposed image, red spots are often used to indicate the Cyanine 5 channel fluoresced more than the Cyanine 3 channel. Green spots indicate the Cyanine 3 channel fluoresced more than the Cyanine 5 channel. Yellow spots indicate equivalent (or nearly equivalent) fluorescence. Brighter spots reveal probes that attracted larger amounts of labeled cDNA. Dimmer (darker) spots indicate smaller amounts of labeled cDNA were bound (see the figure on the following page).

Intensity Image Drawing

Below is a drawing of an artificial superimposed intensity image similar to that output in image construction software. Bright spots indicate a large amount of cDNA hybridization. Dark spots reflect low or no hybridization. If red, yellow, and green colors could be seen, they would indicate *relative* hybridization at each spot.

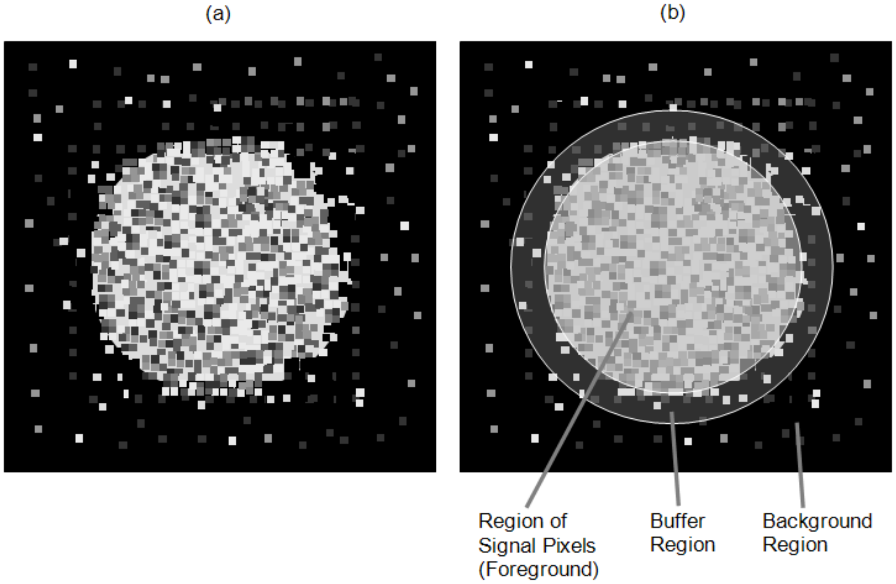


Step 4 – Array Image Processing

There are two main steps in image processing. First, the location of each spot is determined (spot finding or gridding). Second, the pixels of each spot are separated into foreground/signal and background regions (image segmentation). It is intended that the foreground region corresponds to the region where the probe is located, while the background region should be a region where no probe sequences were positioned. Special algorithms, which are continuously being improved, are used in image processing software to perform both of these steps. A properly found and segmented image should look something like (b) in the figure below.

Image Segmentation

The intensity image before spot location and segmentation is seen in (a). The intensity image after spot location and segmentation is shown in (b). A buffer region is used to assure pixels on the border are not misclassified.

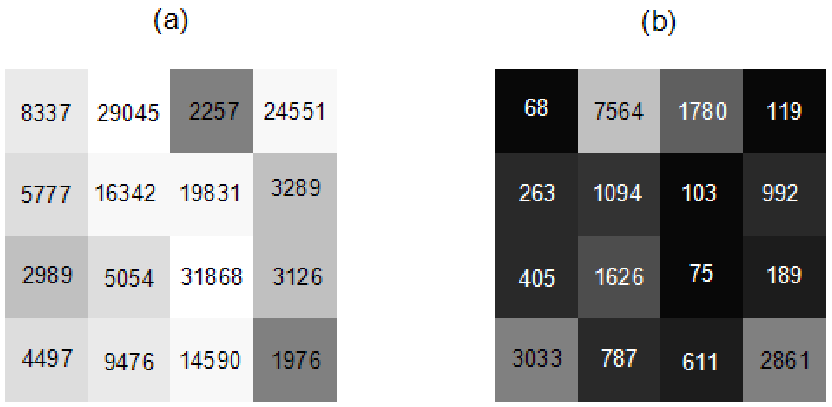


Step 5 – Image Quantification

Summaries of foreground and background pixel regions for each channel that are commonly provided by image analysis software are the mean, median, standard deviation, total intensity, and circularity, among others. The mean or median are usually used as final intensity measures for each spot of each channel. The other measures are used for inspecting spot or array quality or for filtering undesired spots.

Pixel Intensities

The figure that follows shows (a) sixteen pixels from the foreground region of a scanned spot with associated numeric intensities, and (b) sixteen pixels from the background region of a scanned spot with associated numeric intensities. Summaries of these numeric intensities (e.g., mean, median, and standard deviation) are obtained with image analysis software.



(a)

8337	29045	2257	24551
5777	16342	19831	3289
2989	5054	31868	3126
4497	9476	14590	1976

(b)

68	7564	1780	119
263	1094	103	992
405	1626	75	189
3033	787	611	2861

In a reference design, which is described in detail later in the chapter, the two samples to be analyzed in a single array are designated *target* and *reference* samples. The target sample is labeled with either the Cyanine 5 (red) or the Cyanine 3 (green) dye. The reference sample is labeled with the other dye. The goal of image quantification is obtaining a single value that reflects the relative expression of the target to reference channels at each spot. Two common measures of the relative expression of the two channels are the intensity difference (*Target – Reference*), and the log of the intensity ratio, *M*,

$$M = \log_2(Target / Reference) = \log_2(Target) - \log_2(Reference)\,,$$

where *Target* and *Reference* represent, for example, the median foreground intensity for the target and reference channels, respectively. A common measure of overall brightness for each spot is

$$A = \log_2 \sqrt{Target * Reference} = \frac{\log_2(Target) + \log_2(Reference)}{2}$$

M and A are mnemonics for *m*inus and *a*dd (or *a*verage, or *a*bundance), respectively.

Using Logarithms

Dudoit et. al (2002) suggest using logged intensities rather than absolute intensities for the following reasons: "(i) the variation of logged intensities and ratios of intensities is less dependent on absolute magnitude; (ii) normalization is usually additive for logged intensities; (iii) taking logs evens out highly skewed distributions; and (iv) taking logs gives a more realistic sense of variation." They also note that "logarithms base 2 are used instead of natural or decimal logarithms as intensities are typically integers between 0 and $2^{16} – 1$."

Step 6 – Array Spot Types

Although the majority of spots on an array will be used to compare gene expression levels across treatments, some arrays contain spots that are used only for array quality purposes. Below is a description of commonly used types of array quality spots.

Positive (Calibration) Control Spots: Spots containing probes corresponding to genes that are known to be expressed in all cells of the type under consideration. In a two-channel system, it is expected that both channels of positive control spots will have high (well above background) and equal expression.

Negative control spots: Spots containing probes that are known to be *not* expressed in cells of the type under consideration. In a two-channel system, it is expected that both channels of negative control spots will have low (near background) and equal expression.

Spike-in (Ratio) Control Spots: Spots containing probes that hybridize to cDNA that is entered into the cDNA solution in known quantities and proportions. These differ from positive controls in that the cDNA is introduced directly into the hybridizing solution, not expressed in the cells. These spots usually correspond to genes that are known to be *not* expressed in cells of the type under consideration. In a two-channel system, a variety of spike-in control spots are often used to exhibit varying amounts of *relative* expression at varying intensities. For example, one spot may correspond to a 3-to-1 red-to-green ratio, while another spot may designate a 1-to-3 red to green ratio.

Blank (Empty) Spots: Spots that contain no probe sequence and are spotted with water only or with nothing. They are used to estimate background hybridization levels.

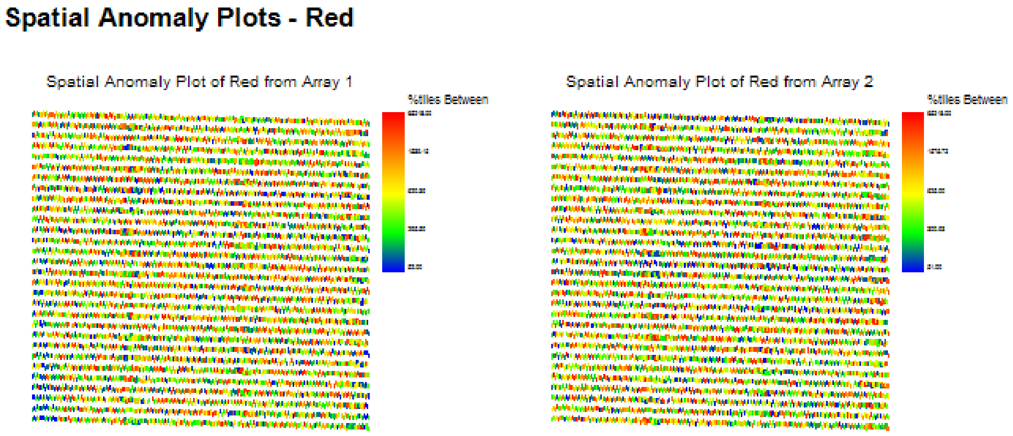
Buffer (Empty) Spots: Spots that contain no probe sequence and are spotted only with buffer solution. They are used to estimate background hybridization levels.

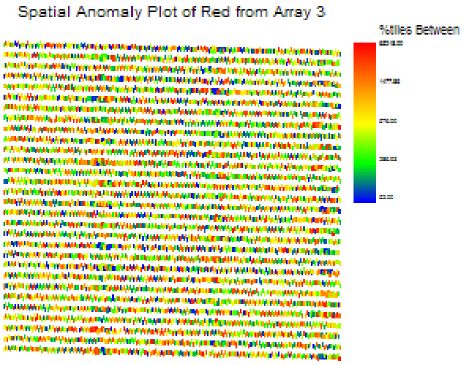
Step 7 – Whole Array Quality

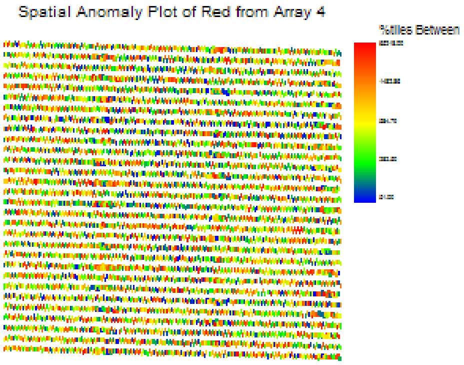
Each microarray should be examined to assure the expression values of the array as a whole can be compared to the corresponding values of other arrays. The following can be used to determine microarrays of questionable quality.

Spatial Anomaly Plot

A spatial anomaly plot is a reconstructed picture of the whole array where intensities are represented by a color spectrum. If there are no anomalies, the distribution of color should be evenly dispersed throughout the plot. The following are four spatial anomaly plots of the median foreground intensity of the Cyanine 5 (Red) channel. There appears to be a region in the lower left area of Array 1 array that shows unusually low intensity.

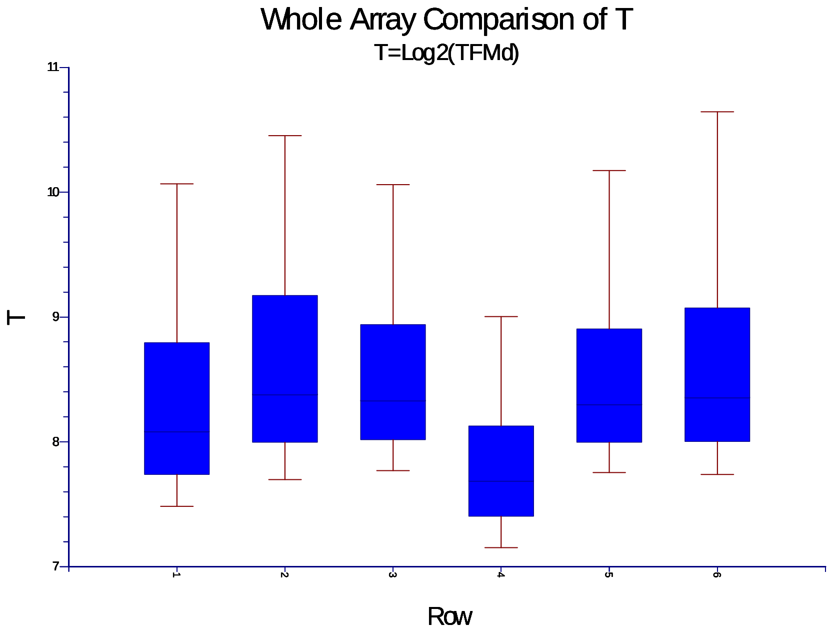






Array Comparison Box Plot

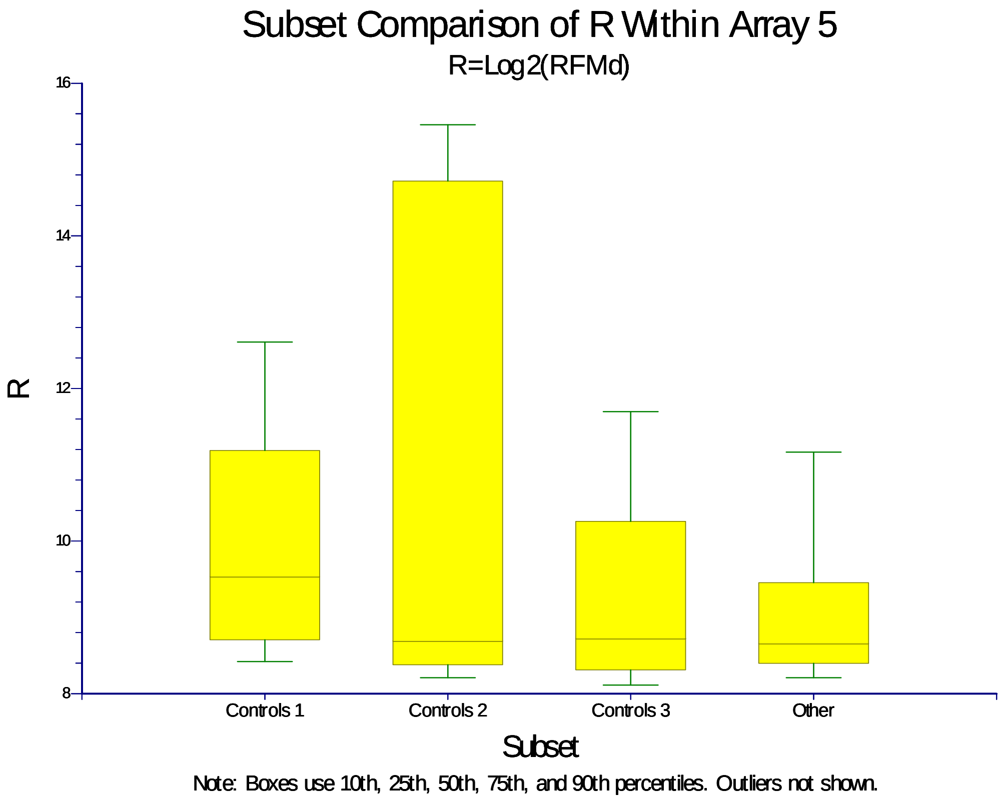
All arrays in the experiment may be compared for a given summary measure using a single side-by-side box plot graph. This example compares the Log2(median foreground intensities) of the Target sample of six arrays.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Subset Box Plot

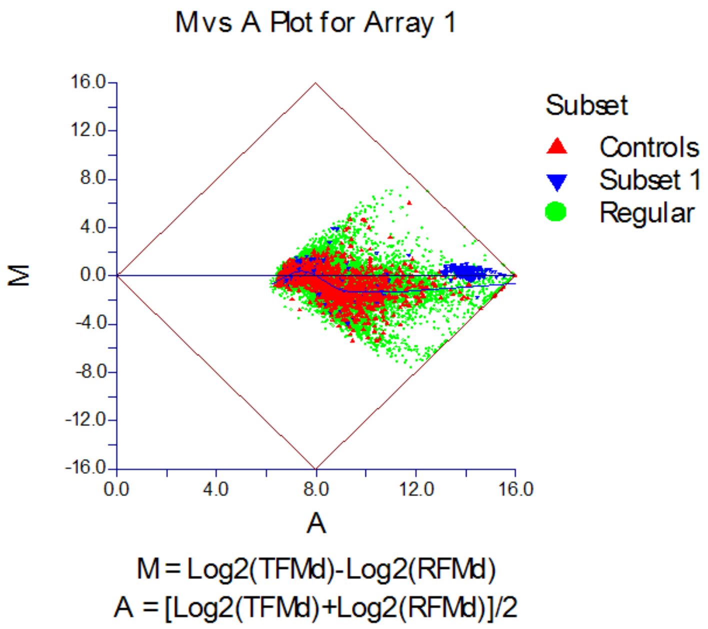
The intensity values of subsets can be compared to other subsets and non-subset spots using a subset side-by-side box plot. Often the subsets are various controls. The subset box plot below shows summaries of the Log2(reference foreground medians) for 3 controls and all other genes (spots).



M vs A Plot

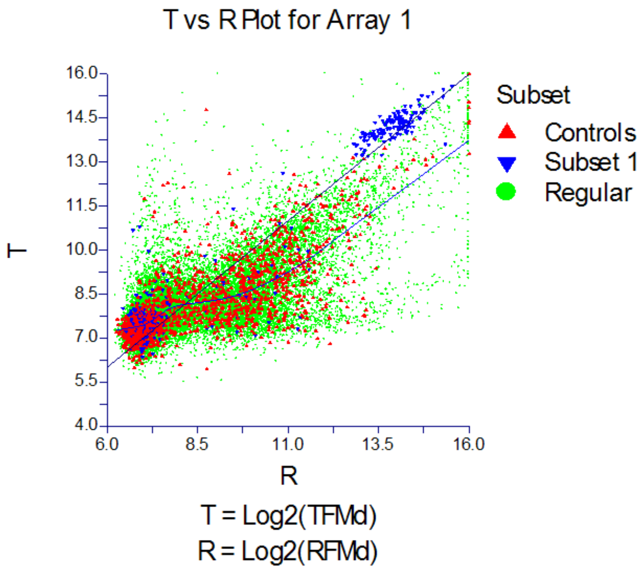
The M vs A plot is a scatterplot with a relative intensity measure (M) on the Y axis, and average intensity (A) on the X axis. This plot is useful for visualizing the relationship between dye-bias and intensity. Dye-bias occurs when one of the dyes produces a brighter image than the other dye for reasons other than differential expression. If there is no dye-bias, M values should be centered near zero across all values of A. If dye-bias exists, a banana shape is often the result. Plotting a loess line on the MA plot can be useful for viewing dye bias. A loess line is a specialized robust moving average that uses locally weighted polynomial regression. If the loess line is far from the zero line, there is evidence of dye bias. If the loess line is near the zero line, little or no dye-bias exists. An M vs A plot following loess normalization can be used to determine if the dye-bias problem has been properly corrected.

If control spots are shown in different colors, the M vs A plot is also useful for displaying how expression levels of control spots compare to those of the spots of interest. The diamond shows the physical limits of the M vs A coordinates when a 16-bit scanner is used. The overlaid loess line shows minor dye bias.



T vs R Plot

The relative intensity values of target and reference samples can be seen by plotting the target intensity measure on the Y axis, and the reference intensity measure on the X axis. This plot is similar to a rotation of the MA plot. This T vs R Plot displays relative median foreground expression of target to reference samples. The upper physical limits of 16 can be seen here.



Numeric Summaries

Mean and standard deviation summaries for mean, median, or standard deviation of foreground and background regions of the spots are useful for comparing slides. Filter summaries across slides may also be good indicators of slides of questionable quality.

Step 8 – Array Individual Spot Quality and Filtering

Some useful indicators of individual spot quality include:

Standard deviation of Foreground or Background Pixels: This summary (of each channel) gives an idea of the uniformity of expression across the spot. Large standard deviations indicate non-uniformity. A large standard deviation can be determined by examining the whole slide numeric summary of standard deviations.

Flags: Some image processing software packages generate flags for spots which fail to meet some criteria. The criteria should be understood before using the flags as indicators of spot quality.

Saturation: Since each pixel has a limiting value of 65,535, pixels with such a value should be considered right-censored (unknown, but larger than 65,535). Spots with large proportions of saturated pixels may be termed saturated spots.

Weak Signal: Spots with low intensities are often near, or even below, the expression intensity of the background region. When expression intensities are near the background, it is difficult or impossible to estimate reliably the true expression level of that gene. It is only known that the gene is not highly or medium expressed. It is not known, however, whether the gene is expressed at very low levels, or not at all. Weak signal spots are essentially left-censored.

Weak Signal Considerations

Another aspect of foreground to background comparison should be mentioned here. Because the background is devoid of probe, it is not subject to nonspecific binding, as is the foreground, and thus may not accurately reflect the baseline it is intended to represent.

Many filters can be designed to remove values of questionable individual spot quality based on some cut-off value. However, it is much easier to flag a spot as one with questionable quality than it is to justify removal of the spot from future analysis. Spot filtering should not be treated lightly, as the following discussion illustrates.

An important decision that may have heavy bearing on the final analysis results is how the spots with low intensities will be treated. When the expression level of either channel is unknown, the relative expression of the two channels is also unknown. For example, suppose that for the Cyanine 3 (Cy3, green) channel at a given spot the estimated foreground is 100 and the estimated background is 150. The true expression intensity could be anywhere from 0 to 150 or more, but the techniques being used are not sensitive enough to distinguish a reliable estimate, due to background noise. Suppose further that the Cyanine 5 (Cy5, red) channel yields estimates of 600 for the foreground and 100 for the background intensity. Because 600 is well above 100, it can be assumed that the Cyanine 5 channel estimate is a reliable one. The difficulty in this situation lies in determining a good estimate of Cyanine 5 to Cyanine 3 relative intensity. The range of ratios is 600/150 = 4 to 600/0 = infinite. Even if logs are used, the ratio chosen will greatly affect the ensuing analysis, as will simply throwing the spot out, altogether.

Consider the following scenario in which the researcher throws out all weak-signal (or low relative intensity) spots. One thousand genes are to be compared based on 10 individuals each from a disease group and a non-disease group. It is of interest to determine which among the 1,000 genes is differentially expressed. Suppose there are four genes, which, unknown to the researcher, are highly expressed in the disease group, but not expressed in the non-disease group. The microarray experiment is run and the expression values for each of those four genes for all 10 individuals in the disease group are well above the background. However, the expression levels for those same four genes for the non-disease group are near background levels. Because the expression levels are near the background, the non-disease expression values for these four genes are all filtered from further analysis. A two-sample *t*-test is then attempted for each of the 1,000 genes. Unfortunately, the *t*-test cannot be run for the four genes of true differential expression, since for each of these four genes there are no expression values in the non-disease group. No evidence of differential expression is reported, and the four genes go undetected.

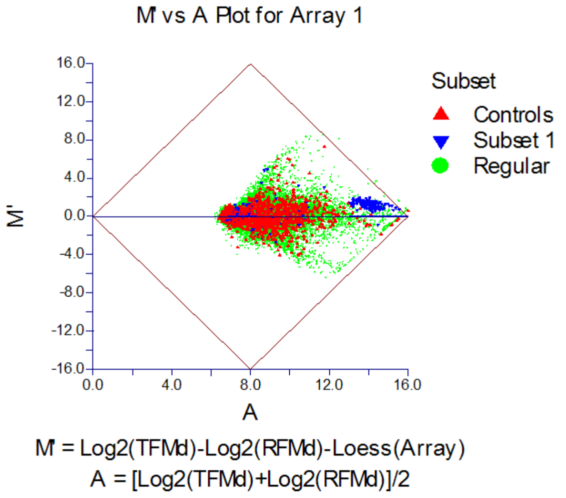
The preceding example reveals three aspects that should be considered when determining whether or not to filter a low intensity spot: (1) The foreground relative to the background for each dye individually; (2) The relative intensities of the two dyes at that spot; (3) The effect of removing that spot from the analysis, *across arrays*.

Step 9 – Array Normalization

The purpose of two-channel array normalization is correcting for dye-bias. Dye-bias occurs when one of the dyes produces a brighter image than the other dye for reasons other than differential expression. Dye-bias is usually dependent on spot intensity and may be viewed using an MA plot with the loess line overlaid (see *MA Plot* under the heading *Two-Channel Whole Array Quality* above). To correct for intensity dependent dye bias at each *A* value, the loess value is subtracted from each *M* value at that location. The *M* values become normalized *M* values. This normalization causes the *M* values to be centered at zero along the horizontal axis (see the figure below) and the dye-bias has been removed.

M vs A Plot Following Array Normalization

Below is the M vs A plot of the same data used for the M vs A plot shown previously, but here following whole array loess normalization. The overlaid loess line follows the zero line, indicating the dye bias has been removed. Some points may go beyond the physical limits of the diamond following loess normalization.



Agilent® Feature Extraction (.txt) Files

Agilent® Feature Extraction is image analysis software. Among its tools is a set of feature-finding algorithms, which are used to find and align each spot on the scanned array. Foreground and background regions are also determined. The foreground and background pixel intensities of each channel are summarized in a file with the suffix .txt. Agilent® scanners and software are compatible with many commercially available slide arrays, including custom lab-printed arrays.

Agilent (.txt) Files

Using Agilent® Feature Extraction software, a .txt file is generated for each array that is scanned. Each .txt file contains a header of several lines followed by the result (feature) columns. The header lines contain information about the scanning, gridding, and settings, as well as summary statistics of all features of the array. The number of lines in the result columns corresponds to the number of spots on the array. The first line of the result (feature) columns contains the column headings. The following is a description of the columns which are commonly used in GESS.

FeatureNum The number associated with the feature.

ProbeName An identifier for the probe, assigned by Agilent®.

GeneName An identifier for the gene that corresponds to the probe.

SystematicName An identifier for the target sequence that corresponds to the probe. A public database identifier is used here if available. When the GeneName is the same as the Systematic name, the Systematic name may not be reported.

PositionX The horizontal coordinate of the center of the spot (in microns).

PositionY The vertical coordinate of the center of the spot (in microns).

rMedianSignal The median foreground pixel intensity for the red channel.

rMeanSignal The mean foreground pixel intensity for the red channel.

rPixSDev The standard deviation of the foreground pixel intensities for the red channel.

rBGMedianSignal The median background pixel intensity for the red channel.

rBGMeanSignal The mean background pixel intensity for the red channel.

rBGPixSDev The standard deviation of the background pixel intensities for the red channel.

gMedianSignal The median foreground pixel intensity for the green channel.

gMeanSignal The mean foreground pixel intensity for the green channel.

gPixSDev The standard deviation of the foreground pixel intensities for the green channel.

gBGMedianSignal The median background pixel intensity for the green channel.

gBGMeanSignal The mean background pixel intensity for the green channel.

gBGPixSDev The standard deviation of the background pixel intensities for the green channel.

rIsSaturated A saturation flag for the red channel for each feature. A value of 1 indicates saturation, while a value of 0 indicates not saturated. A feature is considered saturated if the percent of pixels above the saturation threshold is above 50%.

gIsSaturated A saturation flag for the green channel for each feature. A value of 1 indicates saturation, while a value of 0 indicates not saturated. A feature is considered saturated if the percent of pixels above the saturation threshold is above 50%.

Two-Channel Designs

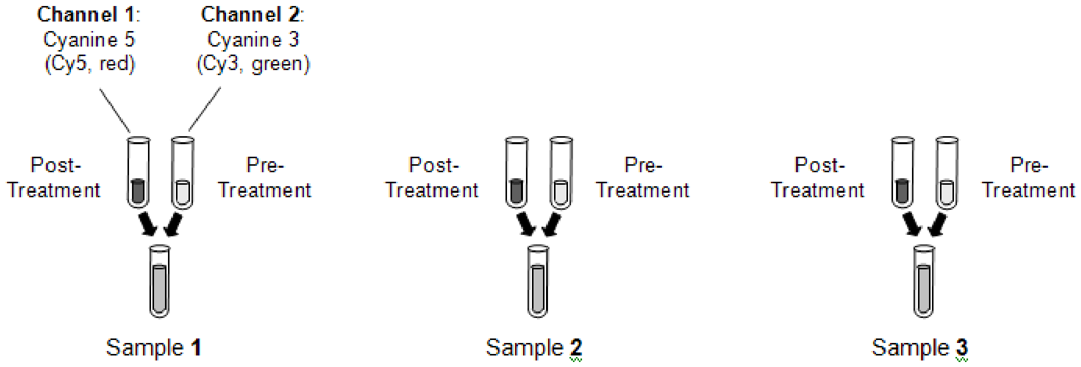
Two experimental designs may be used when using two channel microarrays: paired designs and reference designs.

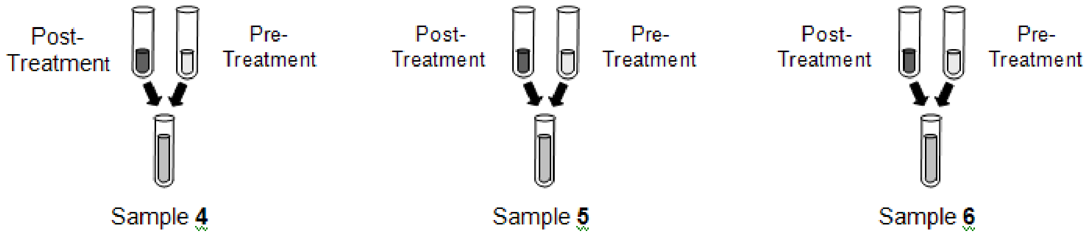
Paired Design

The paired design is often used in two-channel experiments when the gene expression comparison to be made involves a natural pairing of experimental units.

As an example, suppose 6 cell samples are available for comparison. A portion of each of the 6 cell samples (before treatment) is reserved as a control. The same treatment is then given to each of the 6 remaining portions of the samples. It is of interest to determine the genes that are differentially expressed when the treatment is given. In this scenario, there is a natural before/after treatment pairing for each sample. The reserved control portions of each sample are labeled with Cyanine 3 (Cy3, green) dye, while the treatment portions are labeled with Cyanine 5 (Cy5, red) dye. From each sample, the labeled control and the labeled treatment portions are mixed and exposed to an array. The control and treatment portions compete to bind at each spot. The expression of treatment and control samples for each gene is measured with laser scanning. A pre-processing procedure is then used to obtain expression difference values for each gene. In this example, the result is 6 relative expression values (e.g., $\text{Log}_2(\textit{Post} / \textit{Pre})$) for each gene represented on the arrays.

Paired Design, Six Arrays





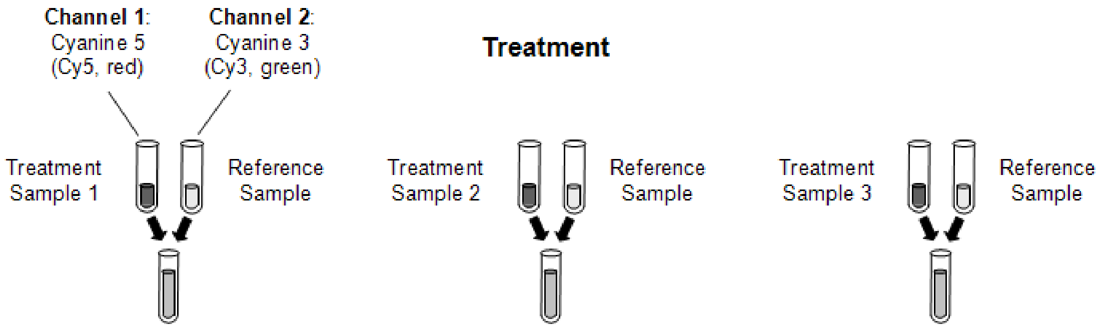
Reference Design

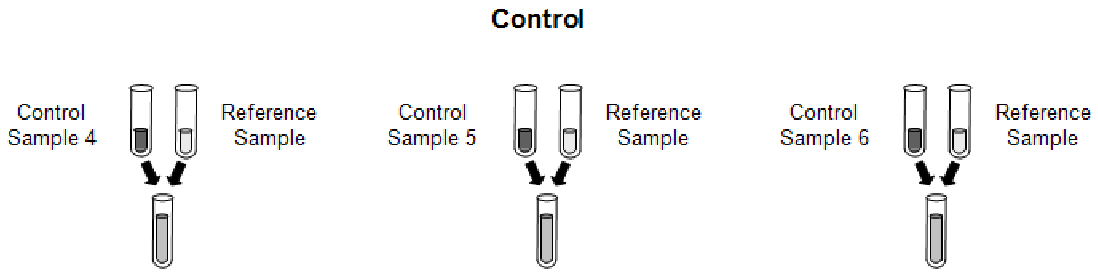
A two-sample reference design, or common reference design, employs an outside source of cDNA that is used as a reference for all samples in the experiment. Reference cDNA may be purchased separately or may be a combination of all cDNAs in the compared samples (The pros and cons of choice of reference cDNA is beyond the scope of this manual).

Suppose a treatment and control are to be compared. One group of experimental units serves as the control group. The other group of experimental units receives the treatment. Following treatment, cDNA is isolated for each of the experimental units. The cDNA for the treatment and control groups may be termed target cDNA. The target cDNA from both groups is labeled with Cyanine 5 (Cy5, red) dye. An outside source of cDNA, with (hopefully) most genes of interest expressed, is labeled with Cyanine 3 (Cy3, green) dye. This cDNA is the common reference, and is used as a baseline for all arrays of both groups. The intensity value for each gene of each array is the relative expression of the target cDNA to the reference cDNA at each spot (see data examples in the tables that follow).

The goal of the reference cDNA is to remove additional variation that may have been introduced in the experimental procedure. Array differences may be particularly pronounced when large periods of time pass between array hybridizations of a single experiment. Reference designs may also be employed in repeated measures/time-course designs.

Two-Sample Reference Design, Six Arrays



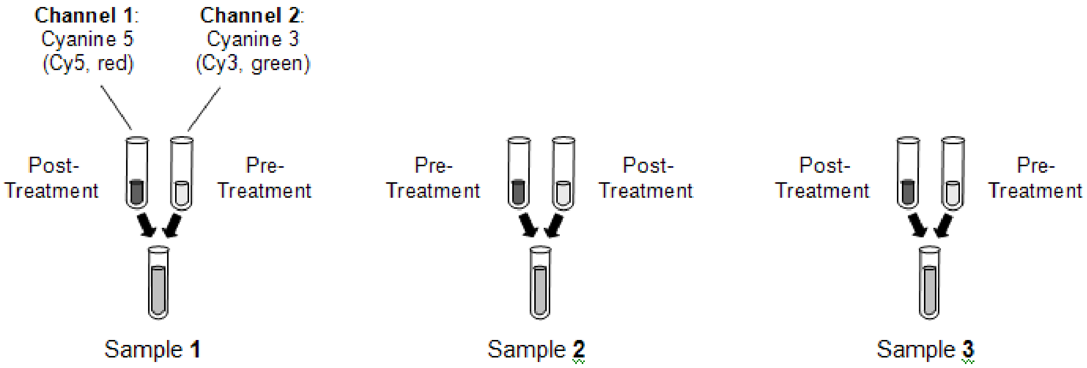


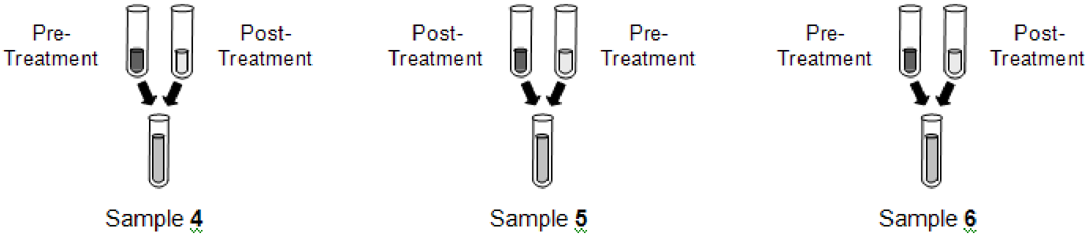
Dye Swap

Dye swap is a technique that may be employed in either paired or reference designs. The purpose is to remove systematic bias of the dye. To use this technique, the dye used is switched for a subset of the experimental units. For example, in the paired design example above, all 6 control portions are labeled with Cyanine 3 (Cy3, green) dye, while the 6 treatment portions are all labeled with Cyanine 5 (Cy5, red) dye. The dye swap technique could be employed by labeling half of the 6 controls with Cyanine 3 (Cy3, green) dye and the other 3 with Cyanine 5 (Cy5, red) dye. The 6 treatment portions would be labeled with the complement dye to that of the corresponding control portions. Careful record should be kept of which dyes are used on each array when performing an experiment with dye-swapping.

Dye-swap - Paired Design

Each of the 6 samples is divided into two portions. One portion serves as control. The other portion receives the treatment. Three of the treatment portions are labeled with Cyanine 5 dye. The corresponding control portions are labeled with Cyanine 3 dye. The other three treatment portions are labeled with Cyanine 3 dye. The corresponding control portions are labeled with Cyanine 5 dye. The samples are combined and then introduced onto a microarray slide. Six relative expression values are obtained for each probe: three are $\mathrm{Log}_2(Cy5 / Cy3)$, the other three are $\mathrm{Log}_2(Cy3 / Cy5)$.





Entering Agilent (.txt) Files

This section describes how file names are entered into the spreadsheet in preparation for preprocessing. Two variables (columns) are required to run Agilent pre-processing, and a third is required to obtain output (.ges) files. Any name (entered by clicking on the Variable Info tab at the bottom of the spreadsheet) may be used for these variables.

Agilent (.txt) File Name Variable

The Agilent (.txt) file name variable is a column on the spreadsheet containing a list of paths and filenames of the .txt files that are to be pre-processed. The files may be in different folders, but must all contain results for the same list of genes. This variable is required to run the Agilent TXT File Pre-Processing Engine procedure.

Target Sample Dye Variable

When two-sample (two-channel) microarrays are used, one sample may be termed the target sample, while the other may be called the reference sample. It is important, particularly when the dye swap technique is used, but also in general, to keep track of whether the target sample is labeled with the red (Cyanine 5, Cy5) dye or the green (Cyanine 3, Cy3) dye. In ***GESS***, this is done with a target sample dye variable. A column is entered into the spreadsheet containing the dye (red or green) of the target sample. Only the values red or green may be entered into this column. This variable is required to run the Agilent TXT File Pre-Processing Engine procedure.

Output File Names Variable

When the Agilent TXT File Pre-Processing Engine is run, a new set of files may be generated for use in statistical analyses. The path and name for these newly created files may be entered into the output file names variable or an empty column may be specified. The files may be in different folders. This variable is required to obtain output for statistical analysis.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options specify the variables that will be used in the analysis.

Agilent Input Files Specifications

These options are used specify the input .txt files that are to be pre-processed.

Agilent (.txt) File Name Variable

Select the variable that contains the list of the Agilent (.txt) array files of the experiment.

The names and pathways of the files should appear in a column below this variable name on the spreadsheet.

Target Sample Dye Variable

Select the variable that identifies whether the target sample is in the red (Cyanine 5, Cy5) channel or the green (Cyanine 3, Cy3) channel.

This required variable must contain either red or green in each cell.

The reference sample is automatically assumed to be in the other channel. That is, if the target sample is in the red (Cyanine 5, Cy5) channel, the reference sample is assumed to be in the green (Cyanine 3, Cy3) channel, and vice versa.

In some experiments, all target samples are in the red (Cyanine 5, Cy5) channel. In other experiments, all reference samples are in the red (Cyanine 5, Cy5) channel. In dye-swap experiments, the target and reference samples are mixed among channels.

In paired sample experiments, the reference and target samples may refer instead to before and after treatment.

Gene Name From

Specify which of the columns (FeatureNum, ProbeName, GeneName, SystematicName) of the .txt file will be used to identify the genes in the output file and ensuing statistical analyses.

Pixel Summary Statistic Settings

These options determine the summary value that will be produced in the output files.

Foreground Statistic

Specify whether the Median, Mean, or Standard Deviation of the pixel intensities is to be used.

For example, if Median is selected here, columns gMedianSignal and rMedianSignal of the .txt file will be used.

NOTE:g is for green; r is for red

Background Statistic

Specify whether the Median, Mean, or Standard Deviation of the pixel intensities is to be used.

For example, if Median is selected here, columns gBGMedianSignal and rBGMedianSignal of the .txt file will be used.

NOTE: g is for green, r is for red, and BG is for background.

Create Target and Reference Using

Specify how the target and reference samples will be summarized.

For example, if 'log2(Foreground - Background)' is specified here, the target samples will be summarized by taking the target foreground statistic, subtracting the target background statistic, followed by taking the logarithm, base2. A similar calculation will occur for the reference.

RECOMMENDATION: We recommend log2(Foreground).

Expression Measure that is Output

The formula selected here indicates the formula that will be used to summarize the expression at each spot of the array. These values are output into a file that can be used in the statistical analyses procedures.

Example: Suppose the target is in the Red channel, 'Median' is selected under 'Foreground Statistic', 'log2(Foreground)' is selected under 'Create Target and Reference Using:', and 'Target - Reference' is selected under 'Expression Measure that is Output:'.

The output file will contain values using the formula log2(rMedianSignal) - log2(gMedianSignal), where r is the red channel, and g is the green channel.

RECOMMENDATION: We recommend Target - Reference - LOESS(Array).

- **Target**

  Only the target sample summary is output. The reference sample is ignored.

- **Reference**

  Only the reference sample summary is output. The target sample is ignored.

- **Target - Reference**

  The reference sample summary is subtracted from the background sample summary.

- **Target – Reference – LOESS(Array)**

  The loess value based on the entire array is subtracted.

GES Output Files Specifications

These options are used to determine the location and naming of the output .ges files.

Folder in which Output Files will be Stored

Enter the path and name of the folder in which the newly created .ges files will be stored. The path may be typed directly or the Browse button may be used to locate the desired folder. New files will be created only if a variable is entered under Output File Names Variable.

In the default path "%mydocs_NCSS%\DATA\GESS", the designator "%mydocs_NCSS%" represents the path to the ***GESS*** personal data folder (commonly *C:\...\[My] Documents\NCSS \NCSS 2007*).

Output File Names Variable

Select the variable in which the path and names of the newly created .ges files will be entered for future statistical analyses. The path and folder of these .ges files is entered under Folder in which Output Files will be Stored.

A new file (to be used for statistical analyses) will be created for each row when the procedure is run.

If this variable is left blank, no new .ges files will be created.

Append to File Names

A sequence of characters may be entered here to append to the name of the created file.

For example if the Agilent .txt file has the name ‘Slide1_10hours.txt’ and ‘log’ is entered here, the newly created .ges file will be ‘Slide1_10hours log.ges’.

If nothing is entered here, the file name will be the same as the name of the .txt file, but ‘.txt’ will be replaced with ‘.ges’.

Overwrite existing output (.ges) files with new output (.ges) files

Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file.

For example, if Slide1.ges has already been created and a new Slide1.ges file is to be written, the new file will be Slide1 (2).ges if the Overwrite box is not checked.

Spreadsheet Filter Active

The filter system allows you to automatically select only certain rows from the database for analysis by this procedure.

The rows to be kept are specified using the Filter option under the Data menu or by pressing the Filter button (the funnel) on the button bar.

USAGE:

If checked, the currently active Filter is used.

If not checked, the filter is not used even if it has been activated in the Filter window.

Reports Tab

The options on this panel control which reports and plots are generated.

Summary Reports

These options are used to determine the reports and report format that are output.

Specification Summary

Check this box to obtain a summary of the formula that is output to the output files, the subsets used, and the filters used.

Array Detail Summary

Check this box to obtain a row by row summary of names of files, numbers of filtered spots, and array means and standard deviations.

Mean Decimals

Specify the number of decimals used for means in the numeric portion of the Array Detail Summary. The number of decimal places is used for display only. It does not change the internal precision of the data.

SD Decimals

Specify the number of decimals used for standard deviations in the numeric portion of the Array Detail Summary. The number of decimal places is used for display only. It does not change the internal precision of the data.

Select Plots

The following options are used to determine which plots will be displayed.

Spatial Anomaly Plot

Check this box to indicate that you want this whole array spatial anomaly plot displayed for all arrays. The spatial anomaly plot gives a spatial view of the entire array for the corresponding measurement. Intensities are separated into four color groupings that reflect four percentile groups. The settings of the spatial anomaly plot are specified under the Spatial Plot tab.

Box Plot - Arrays

Check this box to indicate that you want to display side-by-side box plots comparing all arrays for this measurement. The settings of the box plot comparing array measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Box Plot - Subsets

Check this box to indicate that you want to display side-by-side box plots comparing subsets (control groups) to the primary group of spots for this measurement for all arrays. The settings of the box plot comparing subsets (control groups) measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Scatter Plot – M vs A

Check this box to indicate that you want to display the M vs A plot for each array.

Sometimes called the Ratio-Intensity (R-I) plot, this plot is used to monitor dye bias. The Y axis specifies the value of M, the difference in the intensity summaries of the two samples, Target Summary - Reference Summary. M is a pneumonic for 'minus'. The X axis indicates the value of A, the average intensity summary of the two samples, (Target + Reference)/2. A is a pneumonic for 'add, or average, or abundance'. The loess line is a moving weighted regression average.

Scatter Plot – M’ vs A

Check this box to indicate that you want to display the M' vs A plot for each array.

This plot may be compared to the M vs A plot to see the effect of whole array loess subtraction on dye bias. The Y axis specifies the value of M', where M' = M - whole array loess value. The X axis indicates the value of A, which is the same as in the M vs A plot. The new loess line of the M' values is included.

Scatter Plot – T vs R

Check this box to indicate that you want to display the Target vs Reference plot for each array.

This plot is similar to the M vs A Plot. The Y axis shows the intensity summary of the Target sample. The X axis indicates the intensity summary of the Reference Sample.

Subsets 1 - 9 Tabs

The options on this panel control the names and lists of subsets.

Subset (1 – 9) Name

The name of the gene (spot) subset is entered here.

Plots comparing subsets can be obtained by checking the boxes next to 'Box Plot - Subsets' under the Reports tab. The name chosen here will appear on these plots.

EXAMPLE: To determine whether or not the microarray is functioning properly, it is common to introduce spike-in control DNA into the sample. Spike-in control DNA has a known relative intensity, e.g., 5-fold (target 5 times the reference sample), and corresponds to carefully chosen spots on the array. A list of Spike-in Controls could be given here. Values for the spike-in controls may be compared using box plots to negative controls, positive controls, blank spots, etc. to show that, in fact, the spike-in controls have higher relative intensity values. This would indicate a properly functioning microarray.

Spots in this Subset

Enter a list of genes (spots) that are to be in this subset. The genes (spots) may be entered directly, or the * character may be used to specify all genes with a particular beginning. The gene names or IDs entered in this list must be in the column specified in Spot Name From box on the Variables tab.

EXAMPLES:

Blank

spike1

spike3

spike5

spike*   (all names beginning with spike)

AA44719

NM_00582

NM_04762

NM_27564

cntrl*   (all names beginning with cntrl)

file(C:\Microarray\genelist.txt)   (all names in the genelist.txt file)

var(OutputGenes)   (all names in the spreadsheet variable with the variable name OutputGenes)

Output for Analysis

If this box is checked, the spots in this subset will be included in the output file for future statistical analyses. If this box is not checked, these spots will be removed from the output file.

(Plotting) Symbol

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the plotting symbol. This plotting symbol will be used in the selected array quality graphics.

Filters 1 Tab

The options on this panel control the weak signal filters that will be used.

Filter

Check this box to activate this filter. Spots meeting the criterion will be omitted from the output dataset, but will still be included in pre-processing graphical displays.

Filter Boundary

Specify the filter boundary value. When the spot value is below this boundary value, the output for this spot is omitted from the output dataset. All spots will still be included in pre-processing graphical displays.

Filters 2 Tab

The options on this panel control the saturation, standard deviation, pin group, and negative filters that will be used.

Saturation and SD Filter

Check this box to activate this filter. Spots meeting the criterion will be omitted from the output dataset, but will still be included in pre-processing graphical displays.

SD Filter Boundary

Specify the filter boundary value. When the spot value is above this boundary value, the output for this spot is omitted from the output dataset. All spots will still be included in pre-processing graphical displays.

Spatial Plot Tab

The options on this panel control the features of the spatial anomaly plot.

Heat Map Settings

These settings are used to control the appearance of the heat map and its legend.

Heat Map Colors and Scale

Click on the heat map color bar or the button to the right to change the colors and/or scale of the heat map.

Label

Enter text here for the legend label.

Number of Values

This is the number of reference values printed along the right side of the heat map legend.

Show Legend

Specify whether to show the legend.

Value Format

This option specifies the characteristics of the reference numbers shown next to the heat map legend.

It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot.

It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Plot Settings

These options are used to specify the appearance of the heat map surroundings.

Plot Style file

A plot style file sets all plot options that are not set directly by this procedure.

Interior Color

Specify the interior color of the spatial anomaly plot.

Background Color

Specify the background color of the spatial anomaly plot.

Plotting Symbol Width and Height

This is the width and height in thousandths of an inch of the rectangle that is plotted for each gene.

Recommended:

Width = 120

Height = 150

Plot Titles

Enter text here for the designated title or label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by the long version of the selected intensity summary.

{Y} is replaced by the short version of the selected intensity summary.

{Z} is replaced by the appropriate array number.

Box Plot Tab

The options on this panel control attributes of the box plots.

Vertical and Horizontal Axes

These options control the vertical and horizontal axes attributes.

Axis Labels

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by the appropriate values of the corresponding X axis grouping variable.

{Y} is replaced by the short version of the Y axis intensity summary.

Ref. Number Format

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Major Ticks

Specify the number of large tickmarks and optional grid lines along this axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along this axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along this axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the axis. Thus, checking the Y Grid Lines will actually cause horizontal grid lines to appear.

Box Plot Settings

These options are used to specify the appearance of the box plots.

Box Plot Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created in the Box Plots procedure.

Box Percent Space

When the Box Width (or Bar Width) option is set to Percent Space in the Box Plot Style File selected, this value specifies the percent of the length of the axis that is empty space instead of

bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**

  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**

  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Titles

Enter text for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{G} is replaced by the long version of the Y axis intensity summary.

{Y} is replaced by the short version of the Y axis intensity summary.

{Z} is replace by the appropriate array number.

Box Plot Colors

The options are used to specify the colors of the box plots.

Fill Color

The color used to fill this object. Click to change.

Outline (Border) Color

The color used to outline the object. Click to change.

Line Color

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Scatter Plot Tab

The options on this panel control the main features of the M vs A and T vs R plots.

Vertical and Horizontal Axes

These options control the vertical and horizontal axes attributes.

Ref. Number Format

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Major Ticks

Specify the number of large tickmarks and optional grid lines along this axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along this axis.

Grid Lines

Check this option to display grid lines at the major tickmarks along this axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the axis. Thus, checking the Horizontal Axis Grid Lines will actually cause vertical grid lines to appear.

Scatter Plot Settings

These options are used to specify the appearance of the scatter plots.

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. Scatter plot style files are created in the Scatter Plots procedure.

Symbol

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the points plotted on the scatter plot.

Diamond

Check this box to display a diamond showing the physical boundaries of M on the M vs A, M' vs A, and M" vs A scatter plots.

Zero

Check this box to display a horizontal line at 0 on the M vs A, M' vs A, and M" vs A scatter plots.

45 Degree

Check this box to display a 45 degree line on the T vs R scatter plot only. The 45 degree line shows where T and R are equivalent.

Interior Color

Specify the interior color of the plot.

Background Color

Specify the background color of the plot.

Plot Titles

Enter text here for the designated title or label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{M} is replaced by the long version of the X axis intensity summary.

{S} is replaced by the long version of the Y axis intensity summary.

{X} is replaced by the short version of the X axis intensity summary.

{Y} is replaced by the short version of the Y axis intensity summary.

{Z} is replaced by the appropriate array number.

Loess Options

These options are used to determine whether a loess line is included, its appearance, and the details of how it is computed.

Include Loess Curve

Check this option to display a Loess smooth line.

The locally-weighted, robust regression (loess) smooth is a popular, computer-intensive technique that usually provides a reasonable smoothing of your data without being overly sensitive to outliers. A reasonable smooth is one that travels more or less through the middle of the data. The degree of smoothing is controlled by the Loess % N option.

Loess Order

The order of the polynomial fit in the Loess procedure. Select '1' for a linear fit or '2' for a quadratic fit.

RECOMMENDED: 2 - Quadratic

Loess % N

The percent of the dataset to be used at each Loess calculation.

RECOMMENDED: 40

RANGE: 1 to 99

Number of Points

Specify the number of points at which the Loess line is evaluated. This affects the granularity of the lines. More points imply smoother lines. The number of points selected here may considerably affect the run time.

RANGE: 20 to 2000.

Min-Max Tab

The options on this panel control the minimum and maximum values for the axes of the box plots and scatter plots.

Axis Minimum

Specify the value to be displayed as the minimum on this axis. Data values less than this amount will be ignored. If this value is left blank, the minimum will be determined from the data.

Axis Maximum

Specify the value to be displayed as the maximum on this axis. Data values greater than this amount will be ignored. If this value is left blank, the maximum will be determined from the data.

Labels Tab

The options on this panel control the labels used for scatter plots, spatial anomaly plots, and box plots.

Short Label

Enter here the text that is to be used for the short labels of box plots, spatial anomaly plots, and scatter plots.

Long Label

Enter here the text that is to be used for the long labels of box plots, spatial anomaly plots, and scatter plots. The default is based on the entries under Pixel Summary Statistic Settings of the Variables Tab.

Setup Tab

This panel is used to specify the column headings that are to be read from the Agilent .txt files.

Column Names

An Agilent compatible (.txt) file is made up of many columns. Each column has a corresponding heading name. This program uses the heading name specified here to search for the appropriate column to read from the Agilent (.txt) file. The option to change this name is provided so you can change the column heading name to be read, should the need arise. If you change this heading name to one that does not exist in the file, that column will not be read.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Pre-Processing Agilent TXT Files

This section presents an example of how to pre-process four Agilent .txt files, without involving subsets or filters. The spreadsheet data used are recorded in the AG_Ex1 dataset.

To run this example, take the following steps or load the **Example 1** template from the Agilent TXT File Pre-Processing Engine Template tab.

1 Open the AG_Ex1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **AG_Ex1.S0**.
- Click **Open**.

2 Open the Agilent TXT File Pre-Processing Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Agilent TXT Files**. The Agilent TXT File Pre-Processing Engine procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template. Alternatively, load the Example 1 Template, which generates the specifications described below.

3 Specify the variables.

- On the Agilent TXT File Pre-Processing Engine window, select the **Variables tab**.
- Set the **Agilent (.txt) File Name Variable** to **InputFile**.
- Set the **Target Sample Dye Variable** to **Target**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **OutputFile**.
- Leave **Append to File Names** blank.

4 Specify the Pixel Summary Statistic Settings.

- Continuing on the Variables tab, set the Foreground Statistic to **Median**.
- Set **Create Target and Reference Using** to **log2(Foreground)**.
- Set **Expression Measure that is Output** to **Target – Reference – LOESS(Array)**.

5 Specify the Reports.

- Select the **Reports tab**.
- Check the box next to **Specification Summary** and **Array Detail Summary**.

- Check the **red** and **green boxes** next to **Spatial Anomaly Plot** and **Box Plot – Arrays**.
- Check the **M** and **M'** boxes to the right of **Scatter Plot – vs A**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Expression Formula Output for Analysis

Expression Formula Output for Analysis

Log2(TFMd)-Log2(RFMd)-Loess(Array)

where

Log2: Logarithm Base 2
TF: Target Sample, Foreground Region of Spot
RF: Reference Sample, Foreground Region of Spot
Md: Median Pixel Intensity
Loess(Array): Loess Values Calculated Based on All Values of the Entire Array

This report displays the formula that is used when the output files are created. In this case, the formula may be read as 'log base 2 of the target foreground median minus log base 2 of the reference foreground median minus the whole array loess value.

Value for Spot Was Deleted (Filtered) If

Value for Spot Was Deleted (Filtered) If

No Filters Selected

This report shows that no filters were selected.

Subset Summary

Subset **Subset Values Output for Analysis?**

No subsets were used.

This report shows that no subsets were created.

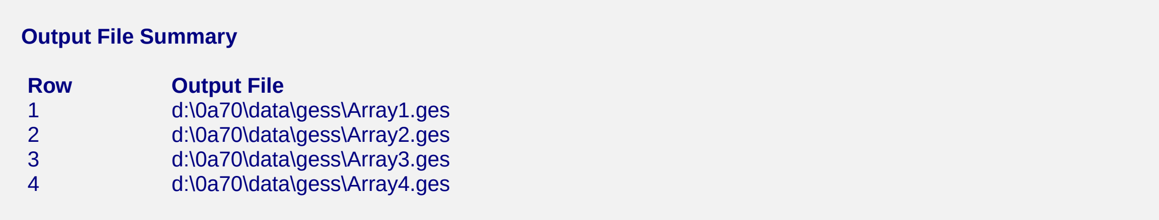
Input File Summary

Input File Summary

Row	Input File
1	…\Data\GESS\AG\Array1.txt
2	…\Data\GESS\AG\Array2.txt
3	…\Data\GESS\AG\Array3.txt
4	…\Data\GESS\AG\Array4.txt

This report shows a list of the input file paths.

Output File Summary

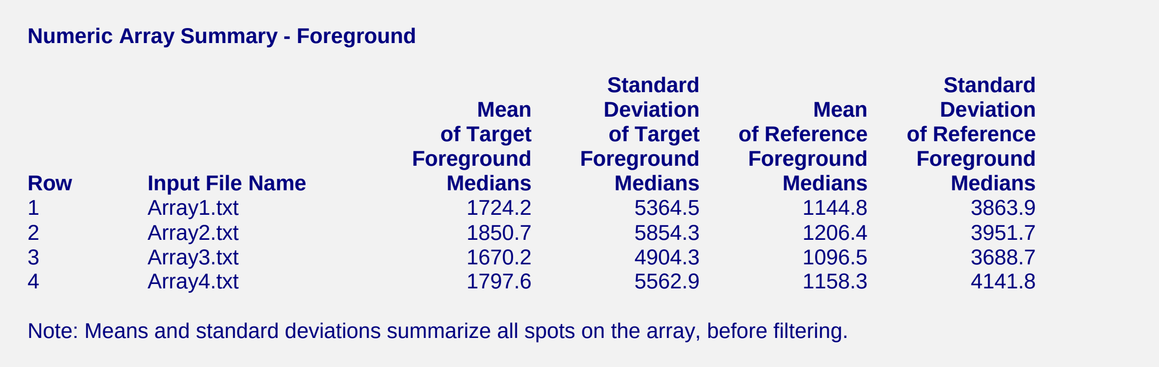


Output File Summary

Row	Output File
1	d:\0a70\data\gess\Array1.ges
2	d:\0a70\data\gess\Array2.ges
3	d:\0a70\data\gess\Array3.ges
4	d:\0a70\data\gess\Array4.ges

This report shows a list of the output file paths. These are the names of the files that will be used as input for statistical analyses.

Numeric Array Summary - Foreground



Numeric Array Summary - Foreground

Row	Input File Name	Mean of Target Foreground Medians	Standard Deviation of Target Foreground Medians	Mean of Reference Foreground Medians	Standard Deviation of Reference Foreground Medians
1	Array1.txt	1724.2	5364.5	1144.8	3863.9
2	Array2.txt	1850.7	5854.3	1206.4	3951.7
3	Array3.txt	1670.2	4904.3	1096.5	3688.7
4	Array4.txt	1797.6	5562.9	1158.3	4141.8

Note: Means and standard deviations summarize all spots on the array, before filtering.

This report shows the whole array foreground region means and standard deviations for target and reference samples.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Mean of Target Foreground Medians

For the target sample, this is the average of all median pixel intensities of the foreground regions of the entire array.

Standard Deviation of Target Foreground Medians

For the target sample, this is the standard deviation of all median pixel intensities of the foreground regions of the entire array.

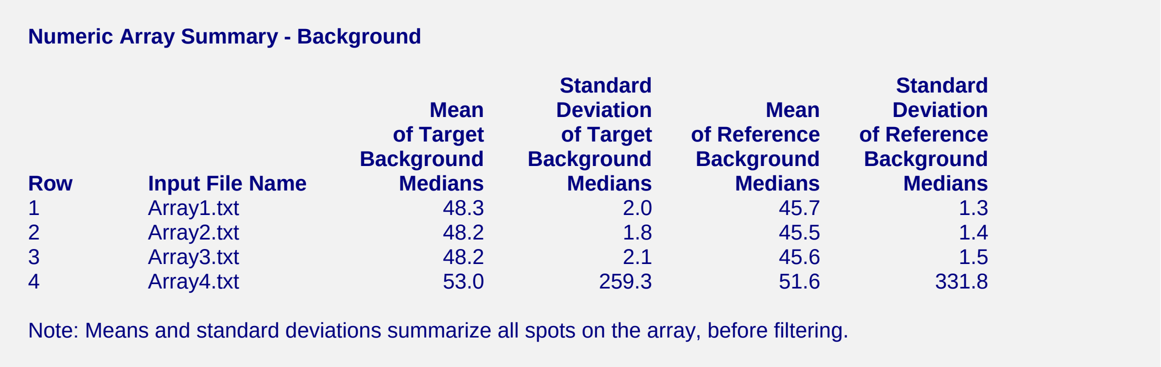
Mean of Reference Foreground Medians

For the reference sample, this is the average of all median pixel intensities of the foreground regions of the entire array.

Standard Deviation of Reference Foreground Medians

For the reference sample, this is the standard deviation of all median pixel intensities of the foreground regions of the entire array.

Numeric Array Summary - Background



Numeric Array Summary - Background

Row	Input File Name	Mean of Target Background Medians	Standard Deviation of Target Background Medians	Mean of Reference Background Medians	Standard Deviation of Reference Background Medians
1	Array1.txt	48.3	2.0	45.7	1.3
2	Array2.txt	48.2	1.8	45.5	1.4
3	Array3.txt	48.2	2.1	45.6	1.5
4	Array4.txt	53.0	259.3	51.6	331.8

Note: Means and standard deviations summarize all spots on the array, before filtering.

This report shows the whole array background region means and standard deviations for target and reference samples.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Mean of Target Background Medians

For the target sample, this is the average of all median pixel intensities of the background regions of the entire array.

Standard Deviation of Target Background Medians

For the target sample, this is the standard deviation of all median pixel intensities of the background regions of the entire array.

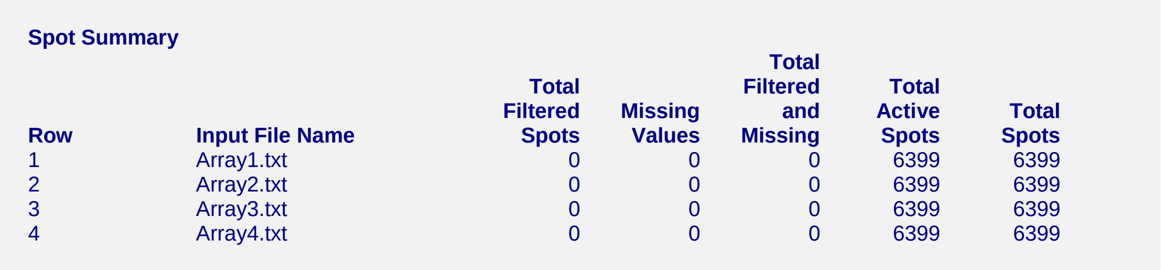
Mean of Reference Background Medians

For the reference sample, this is the average of all median pixel intensities of the background regions of the entire array.

Standard Deviation of Reference Background Medians

For the reference sample, this is the standard deviation of all median pixel intensities of the background regions of the entire array.

Spot Summary



Spot Summary

Row	Input File Name	Total Filtered Spots	Missing Values	Total Filtered and Missing	Total Active Spots	Total Spots
1	Array1.txt	0	0	0	6399	6399
2	Array2.txt	0	0	0	6399	6399
3	Array3.txt	0	0	0	6399	6399
4	Array4.txt	0	0	0	6399	6399

This report shows a summary of filtered spots, missing values, active and total spots.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Total Filtered Spots

This is number of spots that were filtered. The specifics of the filter are found in the next summary.

Missing Values

This is number of missing values among all spots.

Total Filtered and Missing

This is the sum of the Total Filtered Spots and the Missing Values.

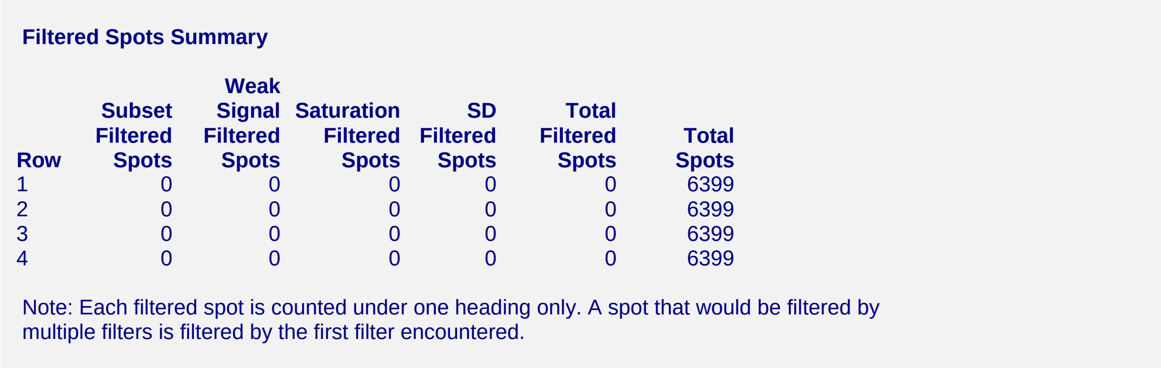
Total Active Spots

This is the number of spots that are not filtered, nor missing.

Total Spots

This is the total number of spots on the array.

Filtered Spots Summary



Filtered Spots Summary

Row	Subset Filtered Spots	Weak Signal Filtered Spots	Saturation Filtered Spots	SD Filtered Spots	Total Filtered Spots	Total Spots
1	0	0	0	0	0	6399
2	0	0	0	0	0	6399
3	0	0	0	0	0	6399
4	0	0	0	0	0	6399

Note: Each filtered spot is counted under one heading only. A spot that would be filtered by multiple filters is filtered by the first filter encountered.

This report shows a detailed summary of all filtered spots.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Subset Filtered Spots

This is the total number of spots that were filtered because they were members of a deleted subset.

Weak Signal Filtered Spots

This is the total number of spots that were filtered based on one or more of the twelve weak signal filters.

Saturation Filtered Spots

This is the total number of spots that were filtered based on one or more of the two saturation filters.

SD Filtered Spots

This is the total number of spots that were filtered based on one or more of the four standard deviation filters.

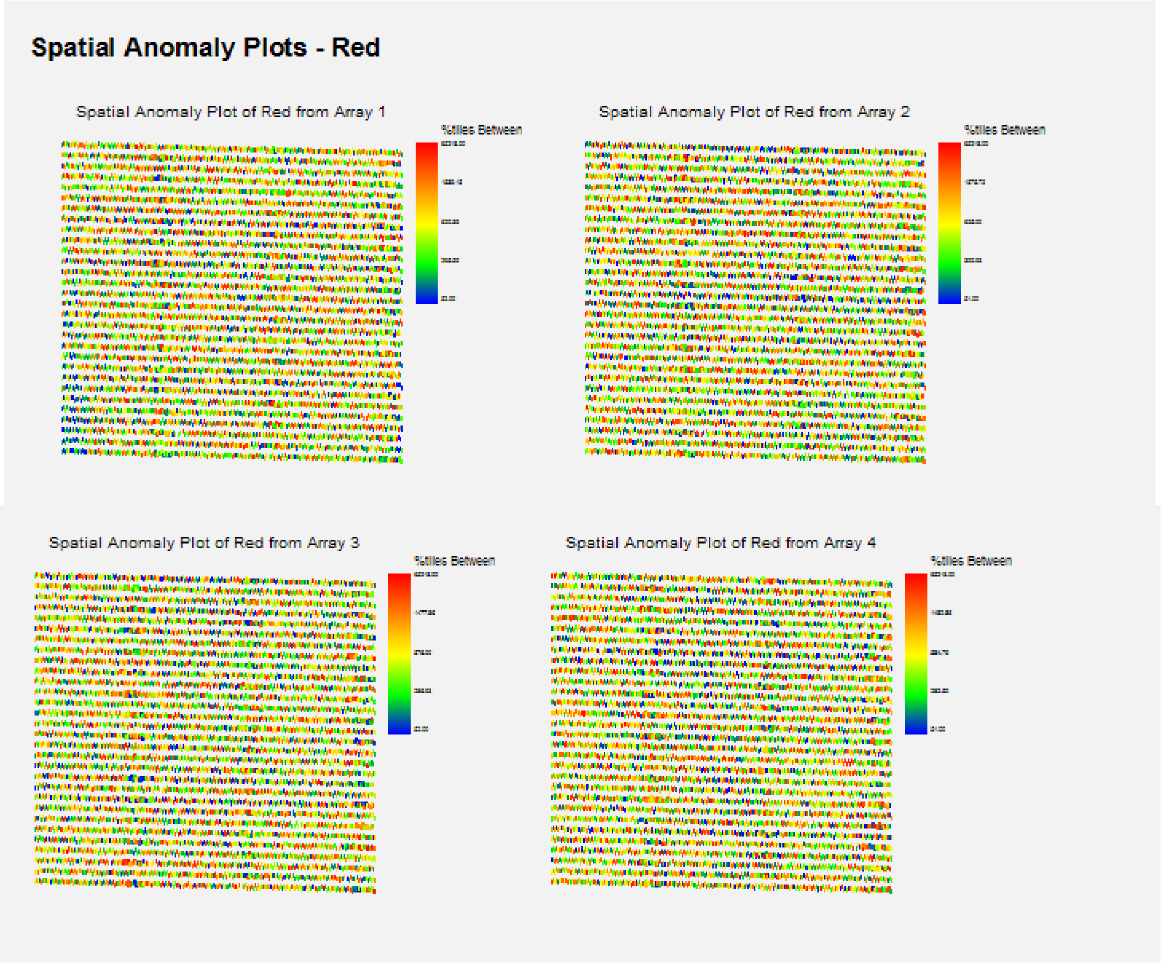
Total Filtered Spots

This is number of spots that were filtered.

Total Spots

This is the total number of spots on the array.

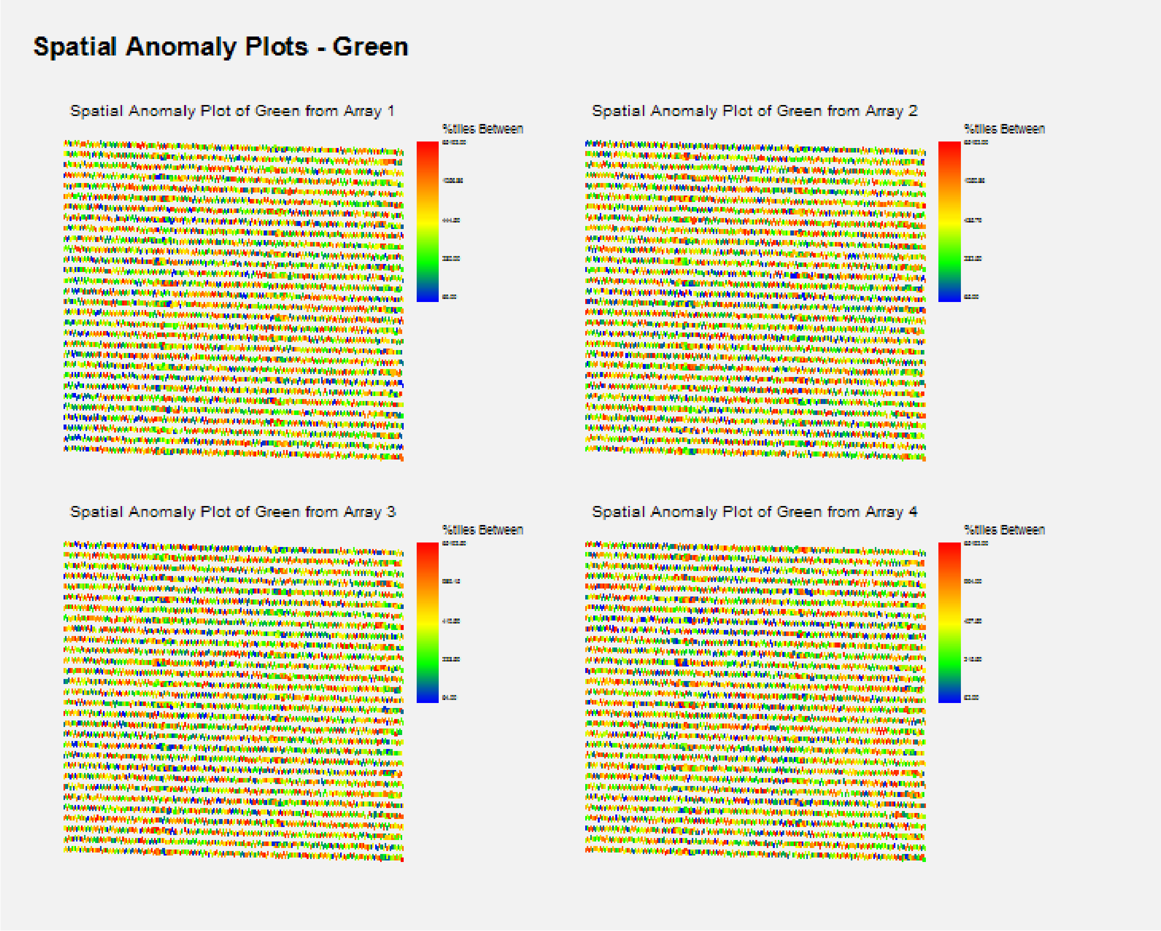
Spatial Anomaly Plots – Red



Spatial Anomaly Plots - Red

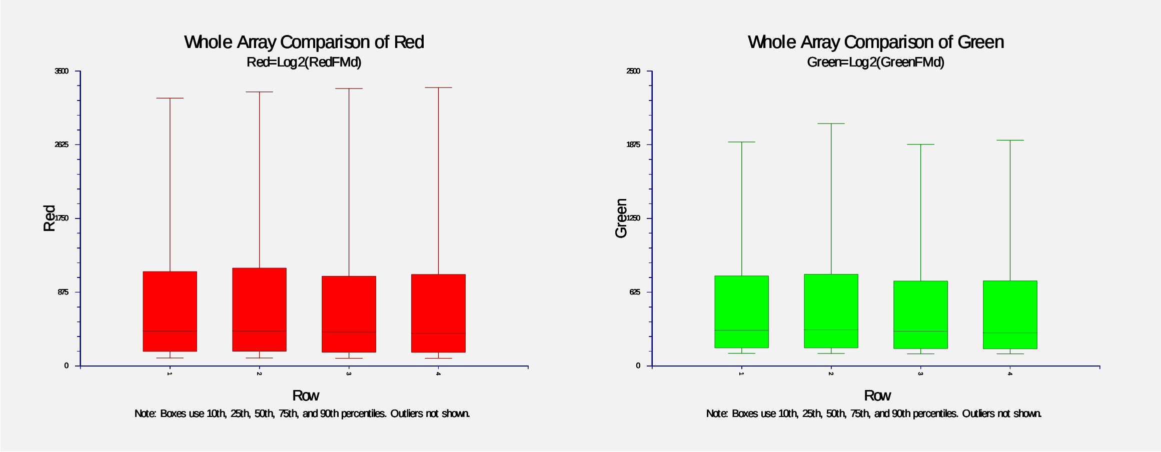
This report shows a spatial representation of the red (Cyanine 5, Cy5) median foreground intensities. There are no definite patterns to indicate spatial problems.

Spatial Anomaly Plots – Green



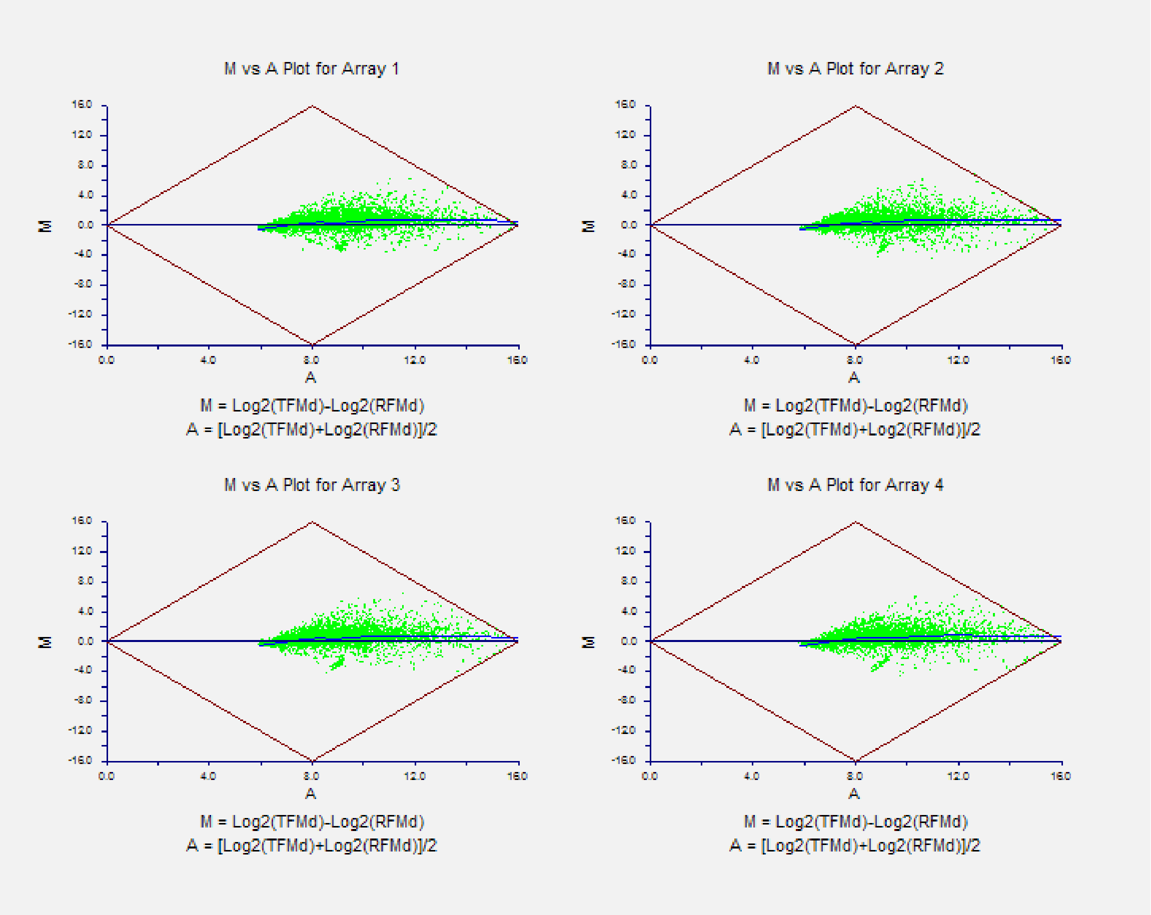
This report shows a spatial representation of the green (Cyanine 3, Cy3) median foreground intensities. There are no patterns to indicate spatial problems.

Array Comparison Section



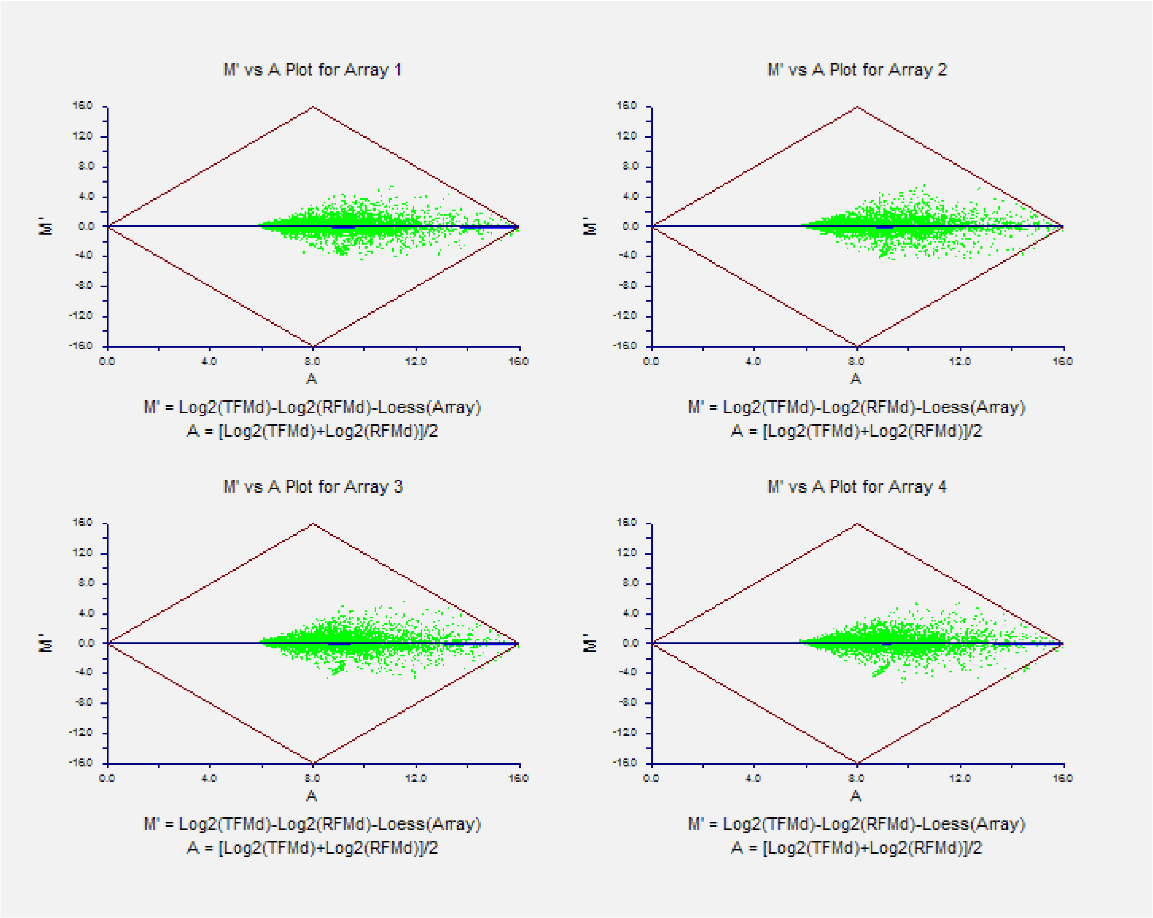
These plots allow comparison of Log2(Foreground Median) of the 4 arrays for both red (Cyanine 5, Cy5) and green (Cyanine 3, Cy3) channels. The overall expression pattern appears similar across arrays.

M vs A Section



These M vs A plots can be used to monitor dye bias. The M value (Y-axis) is the Target minus Reference difference in Log2(Foreground Median) intensities. The A value (X-axis) is the average of the Log2(Foreground Median) intensities. The diamond shows the physical limits of the M vs A coordinates when a 16-bit scanner is used. The loess line is overlaid. If there is no dye bias, the loess line will be near zero.

M’ vs A Section



These M’ vs A plots show the dye bias correction obtained when subtracting the whole array loess value. The loess line is overlaid, but is difficult to see since it is at or near the zero line, indicating the dye bias has been removed.

Chapter 121

GenePix® GPR File Pre-Processing Engine

Introduction

The main purpose of this chapter is to describe the process of obtaining relative expression values from GenePix® output using the ***GESS*** GenePix GPR File Pre-Processing Engine. GenePix Pro® image analysis software is often coupled with the GenePix laser scanners produced by Molecular Devices. GenePix® scanners and software are compatible with many commercially available slide arrays, including Combimatrix®, Amersham®, Illumina®, Agilent®, Schleicher & Scheull®, and custom lab-printed arrays. Following a brief background to the concept of microarrays, this chapter discusses many of the principles and practical aspects of two-channel arrays, including array production, image analysis, array and spot quality, filtering issues, and normalization. Reference and paired experimental designs are also presented as well as the dye-swap technique. The chapter concludes with a tutorial of the entire process of using the GenePix GPR File Pre-Processing Engine to obtain expression values from GenePix® output.

Chapter Structure

Background

An overview of microarray concepts is presented first. This section is designed to familiarize a non-biologist with the concepts of DNA expression and microarray hybridization.

Nine Steps to Obtain Relative Expression Values

The background is followed by a summary of the nine steps required to obtain final relative expression values for a single ***two-channel*** array, which can then be used in comparison analysis.

Step 1 – Spotted Microarray Fabrication. Each probe is spotted on a glass slide by depositing multiple copies of a unique probe sequence. Spots from the same pin are called pin or print-tip groups.

Step 2 – Hybridization. Two samples are each labeled with a different dye, mixed, and then introduced onto the array. The sequences that are complementary to each probe will bind to the probe sequences. The two samples compete for binding at each probe.

Step 3 – Spotted Array Image Construction. The array is scanned twice, once for each dye, producing two image files. The image files contain a value (representing a shade of gray) for every pixel on the array.

Step 4 – Spotted Array Image Processing. Image processing software is used to locate each spot and separate foreground and background regions for both dyes.

Step 5 – Image Quantification. Image processing software produces summaries of the foreground and background pixels for both dyes at each spot. Relative intensities (comparing Cyanine 5 to Cyanine 3 dyes) are also produced.

Step 6 – Spotted Array Spot Types. Results from specially designed probes can be used to assess array quality.

Step 7 – Whole Array Quality. Specialized plots and numeric summaries of the whole array give indication of possible dye bias, print-tip group bias, spatial variation, or other artifacts that indicate whether values obtained from the array can be trusted for comparison.

Step 8 – Spotted Array Individual Spot Quality and Filtering. Pixel summaries and other values for each individual spot give indication of the spot quality. A variety of filters can be used to remove spots of questionable quality.

Step 9 – Array Normalization. A whole array or print-tip group normalization is recommended to correct for dye bias.

The result of these nine steps is a column of relative expression values that can be compared to corresponding values of other arrays in the experiment.

GenePix® Files

The results (.gpr) file and the annotation (.gal) file that are obtained using GenePix Pro® image analysis software are described.

Two-Channel Designs

Paired and Reference Experimental Designs are described, as well as the dye-swap technique.

Entering GPR Files

Details of entering .gpr files into the spreadsheet are explained.

Procedure Options

The options available in ***GESS*** for preprocessing .gpr files are described in detail.

Tutorial/Examples

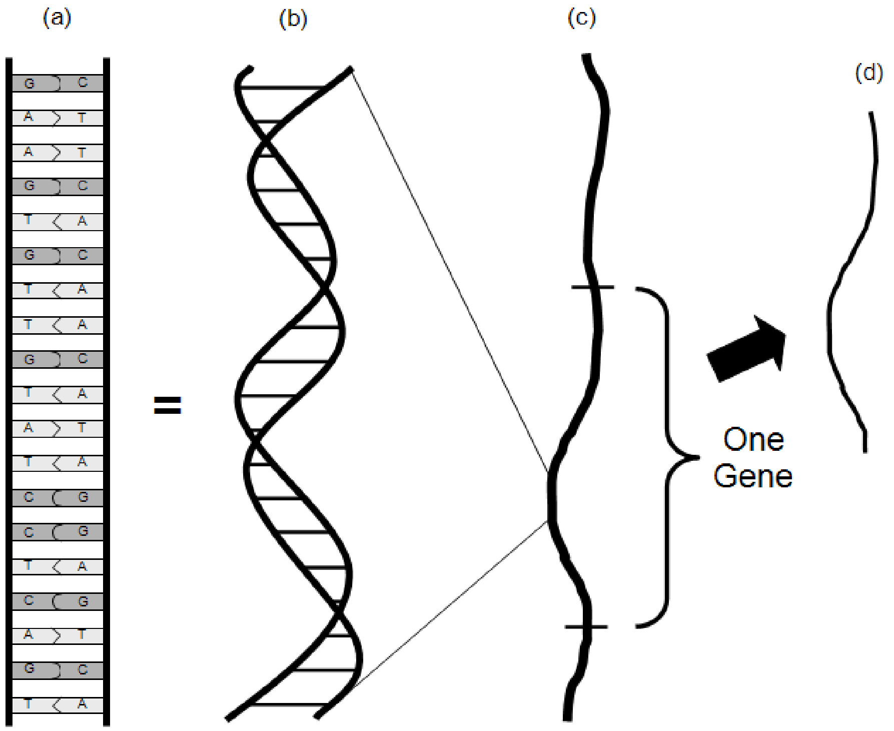
Examples of pre-processing .gpr files in ***GESS*** are shown.

Background

Gene Expression

The general process of gene expression in a cell begins with the DNA. Each DNA molecule has the well-known double helix design. Each rung or step of a DNA molecule is made up of two *nucleotides*. The two nucleotides bonded together are called *base pairs*. In DNA the 4 possible nucleotides are A, T, C, and G (for Adenine, Thymine, Cytosine, and Guanine, respectively). The nucleotide A can only form a base pair with T, and vice versa. Similarly, C can only bind to G, and vice versa. A gene is a unique segment of a DNA molecule consisting of a series of base pairs ranging from about 50 to thousands of base pairs in length. When the need for a specific protein in the cell is identified, the gene for that protein is "read" and a *messenger RNA* (mRNA) is produced in a process called *transcription*. mRNA molecules are single-stranded molecules which are essentially copies of the gene segment of one of the two DNA strands. The mRNA is then used to produce a protein that is specific to that mRNA molecule in a process called translation.

DNA overview. (a) A ladder representation of a DNA segment showing 20 complementary base pairs. (b) A drawing of the three-dimensional form of the corresponding double helix. (c) The 20 base pairs of (a) and (b) are only a small section of the total DNA double strand. A gene is a segment of the DNA helix that contains the code for the production of a protein. (d) A single stranded mRNA molecule is generated when the gene is expressed. The mRNA molecule will be used to produce a protein that is specific to the gene of (c).

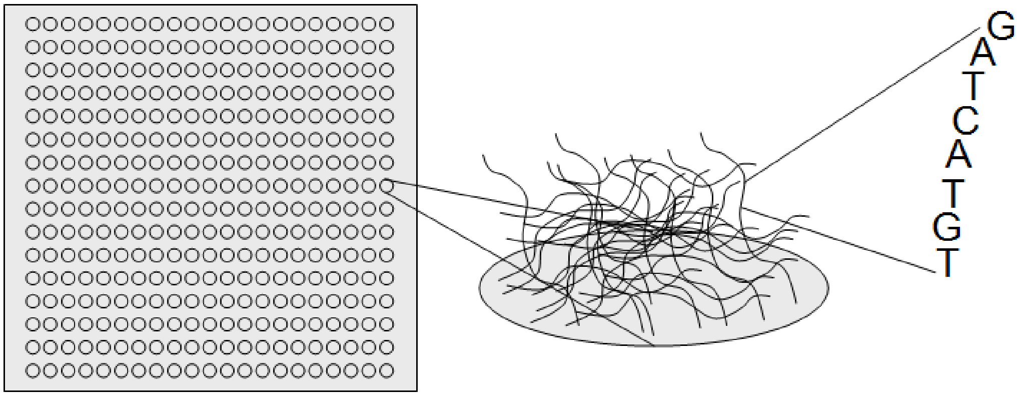


The newly created protein can then be used in the cell to perform the needed function. A gene that is in the process of producing or has produced a protein is said to be *expressed*. Expression of a given gene at a given time can thus be measured either by the amount of mRNA or protein (corresponding to that gene) in the cell. Microarrays are currently the prominent tool for quantifying the amount of mRNA in the cell (or collection of cells) for hundreds or thousands of genes simultaneously.

The Microarray

On a typical glass microarray, there are several thousand spots with *probes* (see below) of known identity, with each probe corresponding to a gene of interest. The probe sequences on each spot are designed to attract only sequences that are expressed by the gene to which that spot corresponds. A spot with the attached probe sequences may collectively be called a probe.

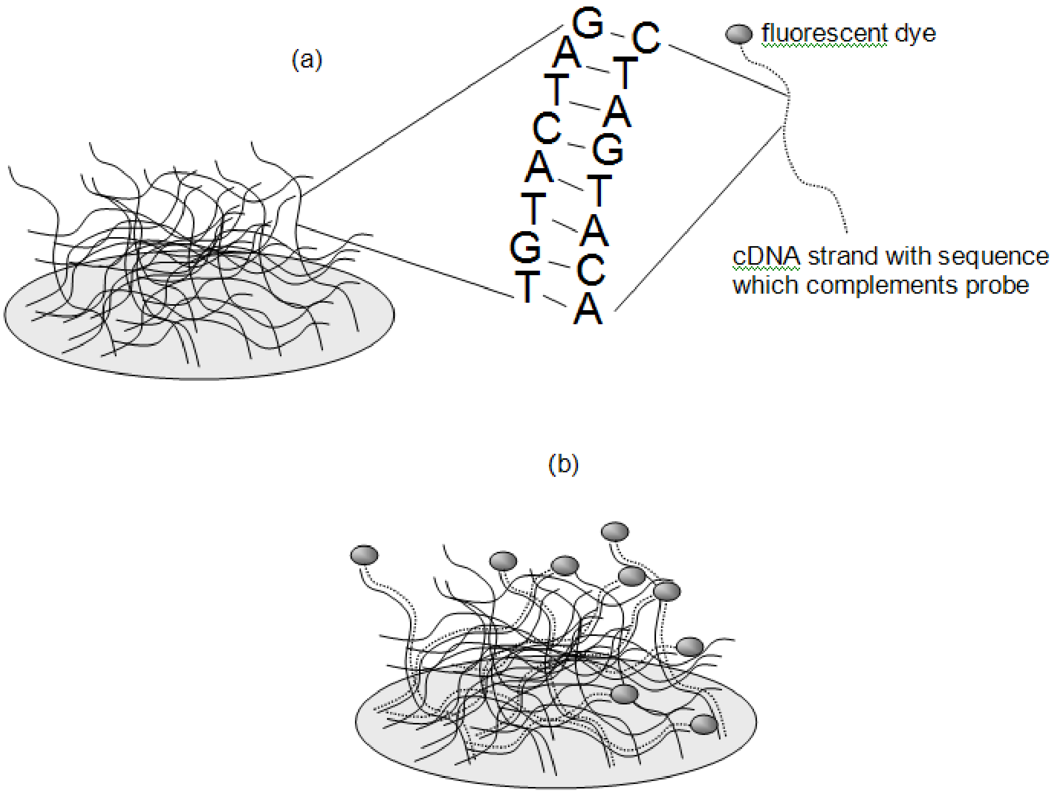
Below is a microarray drawing depicting the arrangement of spots on a spotted array (left), the identical probe sequences on an individual spot (center), and a segment of the probe for this spot that uniquely attracts the mRNA (or cDNA, a more stable mRNA replicate) strands of interest (right). The DNA complement of the probe shown, from bottom to top, is ACATGATC.



Hybridization (Binding to the Microarray)

The mRNA expressed in an experimental unit is obtained from some of the experimental unit’s cells (i.e., blood or tissue), converted to cDNA (a nearly equivalent, but more stable molecule) using a process called reverse transcription, and labeled with fluorescent dye. When the solution containing the cDNA is exposed to a microarray, each of the cDNA sequences will bind to the probe sequences to which it complements. Thus, only sequences with perfect complementation along the entire sequence should bind to the corresponding probe (see figure below). The cDNA from genes which are expressed in higher quantities will hybridize (bind) to the corresponding probe in higher quantities. The amount of hybridized material for each spot can then be measured using the intensity of fluorescence from the bound cDNA when exposed to laser scanning. A scanner (or scanning machine) measures the intensity of fluorescence for every spot on the array. The result of a single microarray scan is several thousand intensities representing the amount of mRNA expression of those genes that are probed on the array.

The figure below shows (a) Only fluorescently labeled cDNA strands that complement the probe along the entire strand will bind to the probe, and (b) A spot with hybridized cDNA, ready for laser scanning.



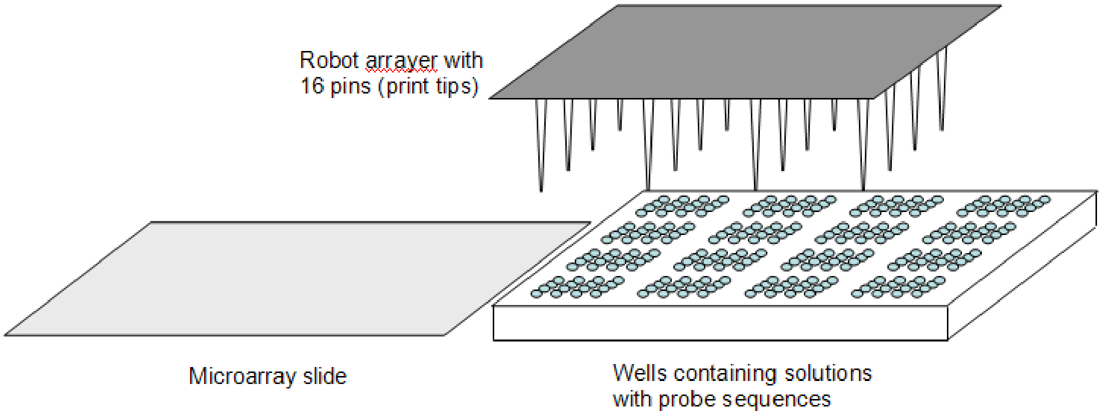
Nine Steps to Obtain Relative Expression Values

Step 1 – Spotted Microarray Fabrication

For spotted (or deposited) arrays, the probe sequences are prepared away from the chip and then spotted onto the slide in small quantities using robots with thin pins. The probe sequence solution for each probe is made by making many copies of a single known sequence using a technique called PCR. The robots dip the pins into each sequence solution and then touch the pins to the surface of the slide to form a spot of probe material (see next page). Groups of spots that are printed with the same pin are called pin groups, or print-tip groups. Because print-tip groups can differ according to the characteristics of each pin, adjustments for differing print-tip groups are often made to expression values in a process called print-tip group normalization.

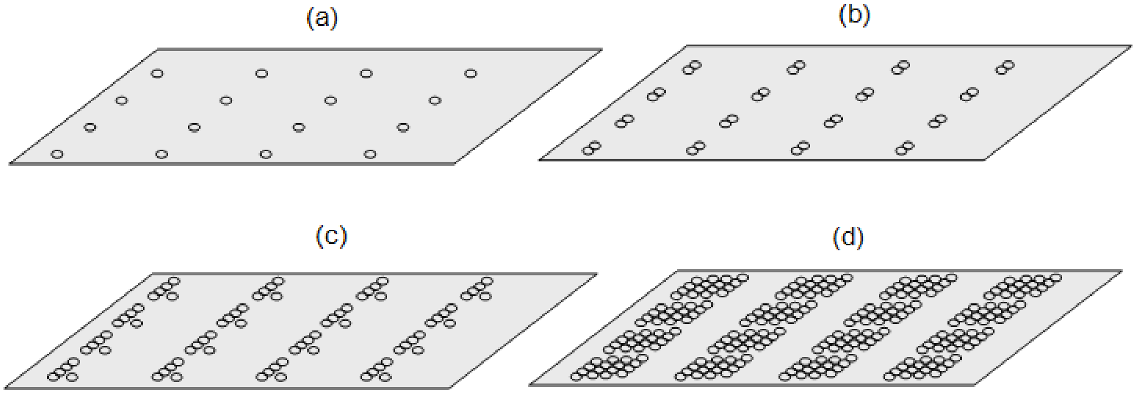
Robot Arrayer

Below is a drawing of a robot arrayer lifting probe solutions to be deposited on the microarray slide. Each well contains a solution with multiple copies of the same sequence. The sequence in each well is different from that of all other wells. The arrayer shifts down one well after each spotting.



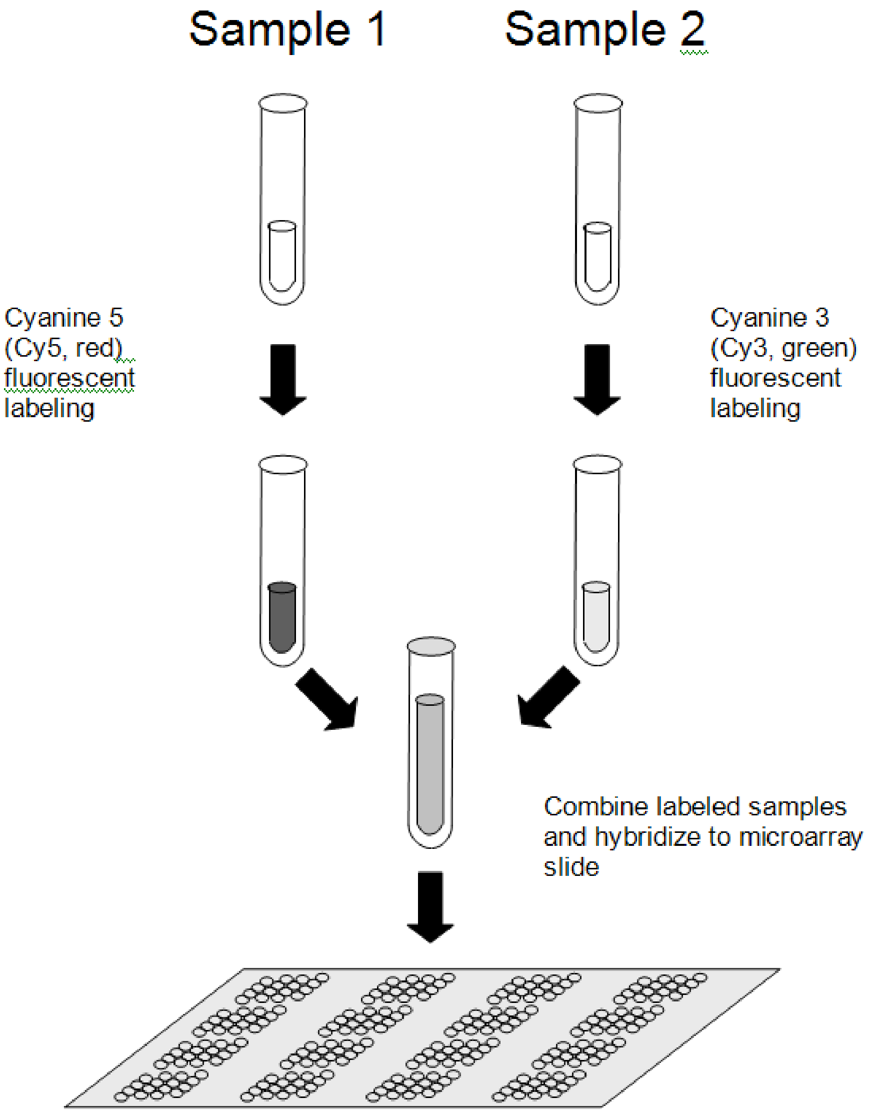
Progress of Fabrication of a Spotted Microarray

The figure below shows (a) the slide after the first group of spots is deposited, (b) the slide after the second spotting, (c) the slide after the fifth spotting, and (d) the slide after the final (16th) spotting. Each block of 16 spots is formed by the same pin and is called a pin group or print-tip group. A total of 16*16 = 256 different probes (representing, perhaps, 256 genes) have been deposited on the final microarray slide.



Step 2 – Hybridization

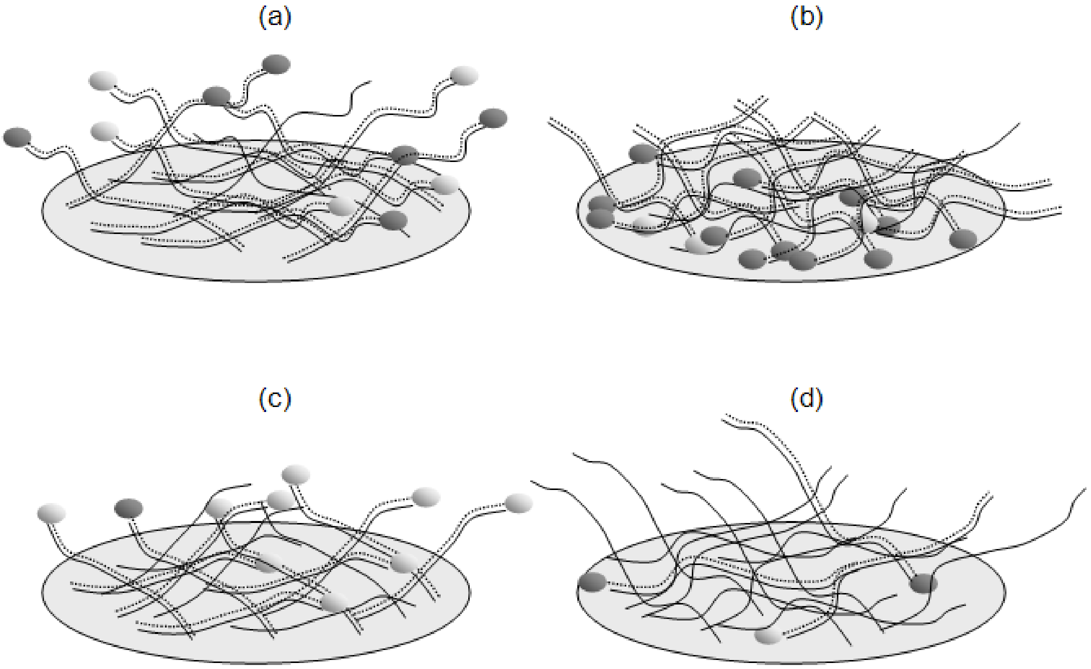
Two-channel microarrays refer to those for which two samples (and two dyes) are analyzed on each array. Two-channel microarrays are also called two-color microarrays. One cDNA sample is labeled with Cyanine 5 (Cy5, red) dye, and the other sample is labeled with Cyanine 3 (Cy3, green) dye. The samples are mixed and then introduced onto the array to compete for hybridization at each spot, as shown in the following diagram of the two-channel fluorescent labeling process.


Sample 1
Sample 2
Cyanine 5 (Cy5, red) fluorescent labeling
Cyanine 3 (Cy3, green) fluorescent labeling
Combine labeled samples and hybridize to microarray slide

If a specific sequence is highly expressed in one of the samples (say, Sample 1) and has low expression in the other sample (say, Sample 2), the probe for that sequence should bind more Sample 1 sequences than Sample 2 sequences. This process is called competitive hybridization.

Competitive Hybridization

Below are some examples of competitive hybridization at individual spots. The dotted line sequences with dark dots attached represent Sample 1 cDNA with Cyanine 5 (Cy5, red) fluorescent labels. The dotted line sequences with light dots attached represent Sample 2 cDNA with Cyanine 3 (Cy3, green) fluorescent labels. Each of the four examples show varying amounts of competitive hybridization. For the probe in (a), there is nearly equal expression among both channels. In (b), there is high expression of this gene in Sample 1, low expression in Sample 2. In (c) can be seen very low expression of this gene in Sample 1, but high expression in Sample 2. There is very low expression for this gene in both samples in (d).

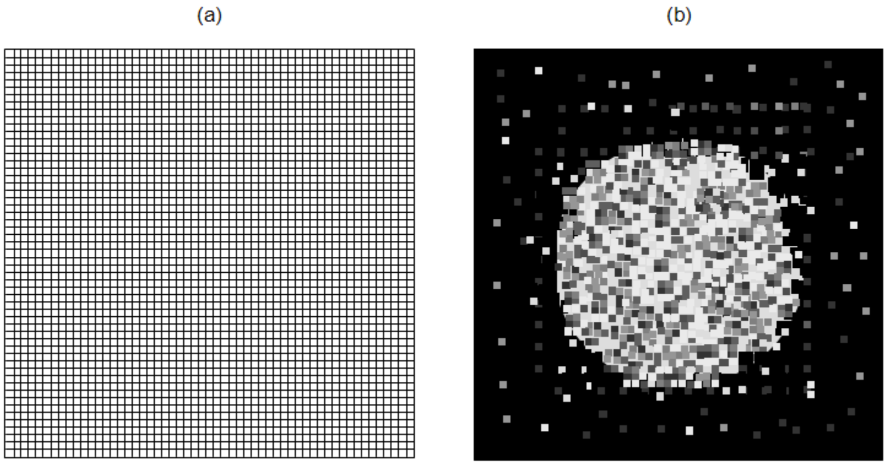


Step 3 – Spotted Array Image Construction

Following competitive hybridization, the laser of a scanning machine is used to illuminate the fluorescent dye of one of the channels (e.g., Cyanine 5) across the whole array, creating a high resolution black-and-white image for that channel. The frequency of the laser is then adjusted (or a different laser is used) to illuminate the fluorescent dye of the other channel (e.g., Cyanine 3), creating a second black-and-white image. Each of the images is usually stored as Tag Image File Format (.tif) file. The image is made up of a grid of pixels. Because each pixel is stored using 16 bits of memory, each pixel can take on any of 2^{16} = 65,536 shades of gray. The numeric range for each pixel is thus 0 to 65,535. The number of pixels in each spot depends upon the resolution (total number of pixels) of the image and the size of the spot (see the figure below).

Pixel Grid and .tiff Image following Laser Scanning

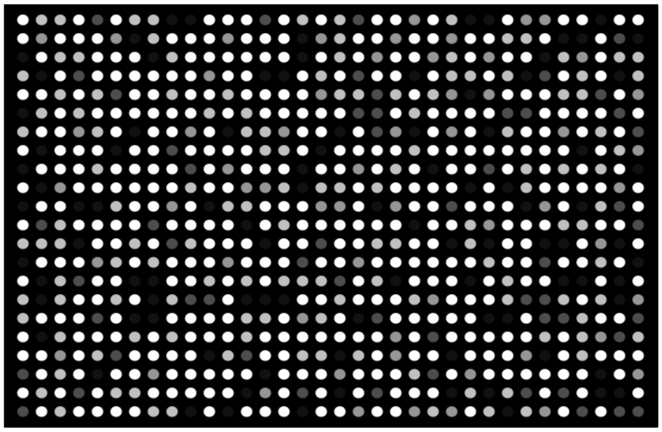
A pixel grid for the region of a single probe spot is shown in (a). A drawing of a .tiff image following laser scanning is shown in (b). The number of pixels in a spot may range from below 50 to several hundred, depending upon the size of the spot and the resolution of the image.



Artificial colors (i.e., red and green) are often used in image construction software to visualize the expression represented on each array. The colors may be superimposed to form a single array image that displays the relative illumination at each spot. On the superimposed image, red spots are often used to indicate the Cyanine 5 channel fluoresced more than the Cyanine 3 channel. Green spots indicate the Cyanine 3 channel fluoresced more than the Cyanine 5 channel. Yellow spots indicate equivalent (or nearly equivalent) fluorescence. Brighter spots reveal probes that attracted larger amounts of labeled cDNA. Dimmer (darker) spots indicate smaller amounts of labeled cDNA were bound (see the figure on the following page).

Intensity Image Drawing

Below is a drawing of an artificial superimposed intensity image similar to that output in image construction software. Bright spots indicate a large amount of cDNA hybridization. Dark spots reflect low or no hybridization. If red, yellow, and green colors could be seen, they would indicate *relative* hybridization at each spot.

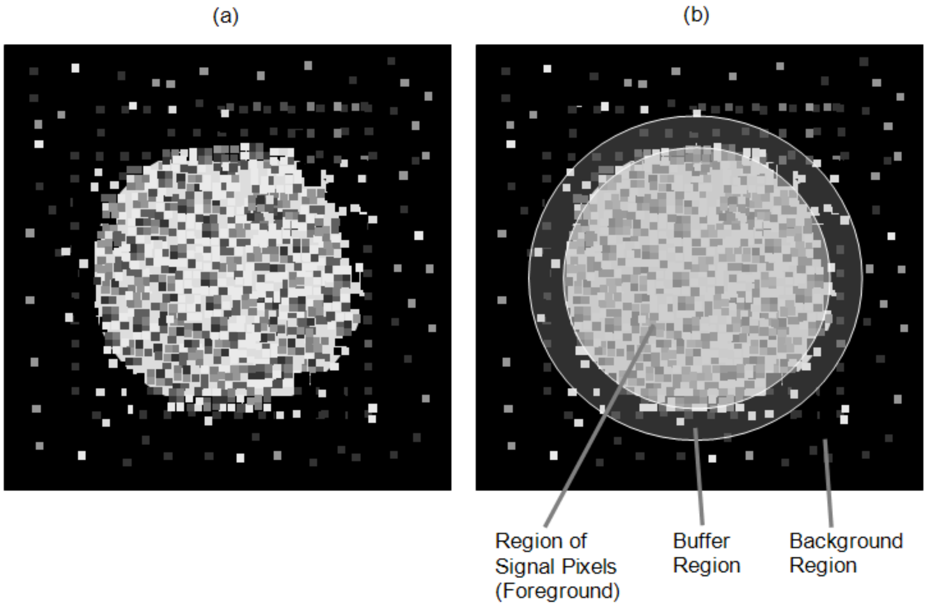


Step 4 – Spotted Array Image Processing

There are two main steps in image processing. First, the location of each spot is determined (spot finding or gridding). Second, the pixels of each spot are separated into foreground/signal and background regions (image segmentation). It is intended that the foreground region corresponds to the region where the probe is located, while the background region should be a region where no probe sequences were positioned. Special algorithms, which are continuously being improved, are used in image processing software to perform both of these steps. A properly found and segmented image should look something like (b) in the figure below.

Image Segmentation

The intensity image before spot location and segmentation is seen in (a). The intensity image after spot location and segmentation is shown in (b). A buffer region is used to assure pixels on the border are not misclassified.

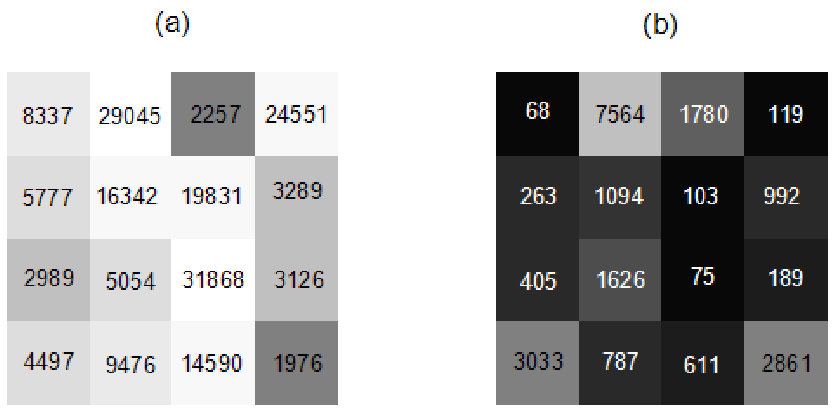


Step 5 – Image Quantification

Summaries of foreground and background pixel regions for each channel that are commonly provided by image analysis software are the mean, median, standard deviation, total intensity, and circularity, among others. The mean or median are usually used as final intensity measures for each spot of each channel. The other measures are used for inspecting spot or array quality or for filtering undesired spots.

Pixel Intensities

The figure that follows shows (a) sixteen pixels from the foreground region of a scanned spot with associated numeric intensities, and (b) sixteen pixels from the background region of a scanned spot with associated numeric intensities. Summaries of these numeric intensities (e.g., mean, median, and standard deviation) are obtained with image analysis software.



(a)

8337	29045	2257	24551
5777	16342	19831	3289
2989	5054	31868	3126
4497	9476	14590	1976

(b)

68	7564	1780	119
263	1094	103	992
405	1626	75	189
3033	787	611	2861

In a reference design, which is described in detail later in the chapter, the two samples to be analyzed in a single array are designated *target* and *reference* samples. The target sample is labeled with either the Cyanine 5 (red) or the Cyanine 3 (green) dye. The reference sample is labeled with the other dye. The goal of image quantification is obtaining a single value that reflects the relative expression of the target to reference channels at each spot. Two common measures of the relative expression of the two channels are the intensity difference (*Target – Reference*), and the log of the intensity ratio, *M*,

$$M = \log_2(Target / Reference) = \log_2(Target) - \log_2(Reference),$$

where *Target* and *Reference* represent, for example, the median foreground intensity for the target and reference channels, respectively. A common measure of overall brightness for each spot is

$$A = \log_2 \sqrt{Target * Reference} = \frac{\log_2(Target) + \log_2(Reference)}{2}$$

M and A are mnemonics for *m*inus and *a*dd (or *a*verage, or *a*bundance), respectively.

Using Logarithms

Dudoit et. al (2002) suggest using logged intensities rather than absolute intensities for the following reasons: “(i) the variation of logged intensities and ratios of intensities is less dependent on absolute magnitude; (ii) normalization is usually additive for logged intensities; (iii) taking logs evens out highly skewed distributions; and (iv) taking logs gives a more realistic sense of variation.” They also note that “logarithms base 2 are used instead of natural or decimal logarithms as intensities are typically integers between 0 and $2^{16} – 1$.”

Step 6 – Spotted Array Spot Types

Although the majority of spots on an array will be used to compare gene expression levels across treatments, some arrays contain spots that are used only for array quality purposes. Below is a description of commonly used types of array quality spots.

Positive (Calibration) Control Spots: Spots containing probes corresponding to genes that are known to be expressed in all cells of the type under consideration. In a two-channel system, it is expected that both channels of positive control spots will have high (well above background) and equal expression.

Negative control spots: Spots containing probes that are known to be *not* expressed in cells of the type under consideration. In a two-channel system, it is expected that both channels of negative control spots will have low (near background) and equal expression.

Spike-in (Ratio) Control Spots: Spots containing probes that hybridize to cDNA that is entered into the cDNA solution in known quantities and proportions. These differ from positive controls in that the cDNA is introduced directly into the hybridizing solution, not expressed in the cells. These spots usually correspond to genes that are known to be *not* expressed in cells of the type under consideration. In a two-channel system, a variety of spike-in control spots are often used to exhibit varying amounts of *relative* expression at varying intensities. For example, one spot may correspond to a 3-to-1 red-to-green ratio, while another spot may designate a 1-to-3 red to green ratio.

Blank (Empty) Spots: Spots that contain no probe sequence and are spotted with water only or with nothing. They are used to estimate background hybridization levels.

Buffer (Empty) Spots: Spots that contain no probe sequence and are spotted only with buffer solution. They are used to estimate background hybridization levels.

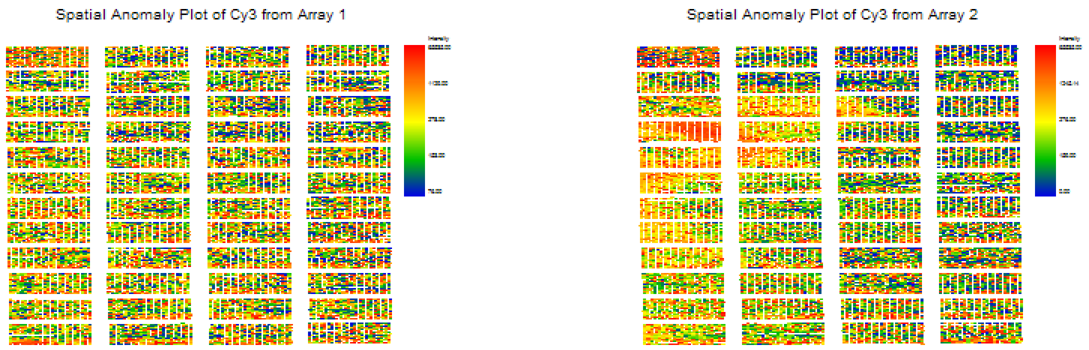
Step 7 – Whole Array Quality

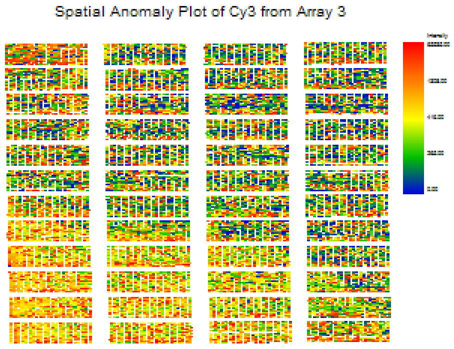
Each microarray should be examined to assure the expression values of the array as a whole can be compared to the corresponding values of other arrays. The following can be used to determine microarrays of questionable quality.

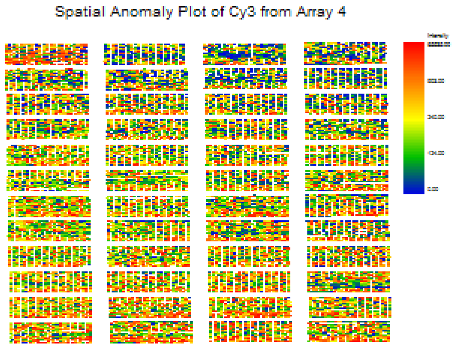
Spatial Anomaly Plot

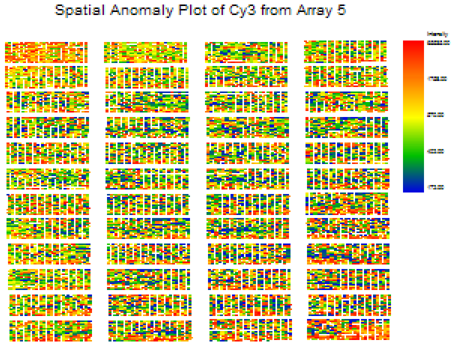
A spatial anomaly plot is a reconstructed picture of the whole array where intensities are represented by a color spectrum. If there are no anomalies, the distribution of color should be evenly dispersed throughout the plot. The following are six spatial anomaly plots of the median foreground intensity of the Cyanine 3 (Green) channel. There appears to be a region in the upper left area of Array 2 array that shows unusual brightness. The lower left region of Array 3 appears brighter than the rest of the array.

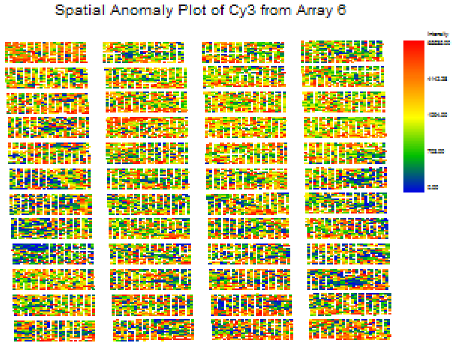
Spatial Anomaly Plots - Cy3




Spatial Anomaly Plot of Cy3 from Array 3

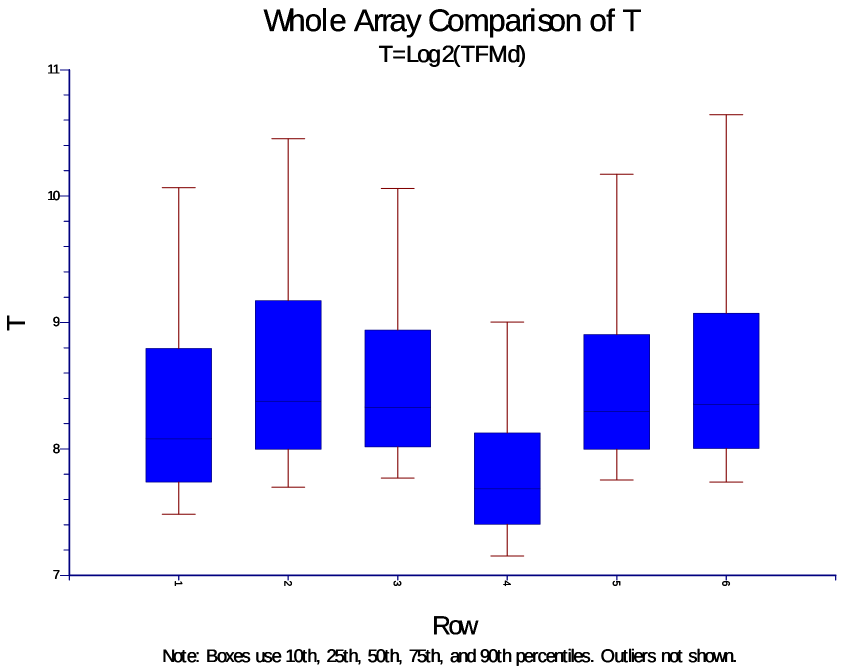

Spatial Anomaly Plot of Cy3 from Array 4


Spatial Anomaly Plot of Cy3 from Array 5


Spatial Anomaly Plot of Cy3 from Array 6

Array Comparison Box Plot

All arrays in the experiment may be compared for a given summary measure using a single side-by-side box plot graph. This example compares the Log2(median foreground intensities) of the Target sample of six arrays.

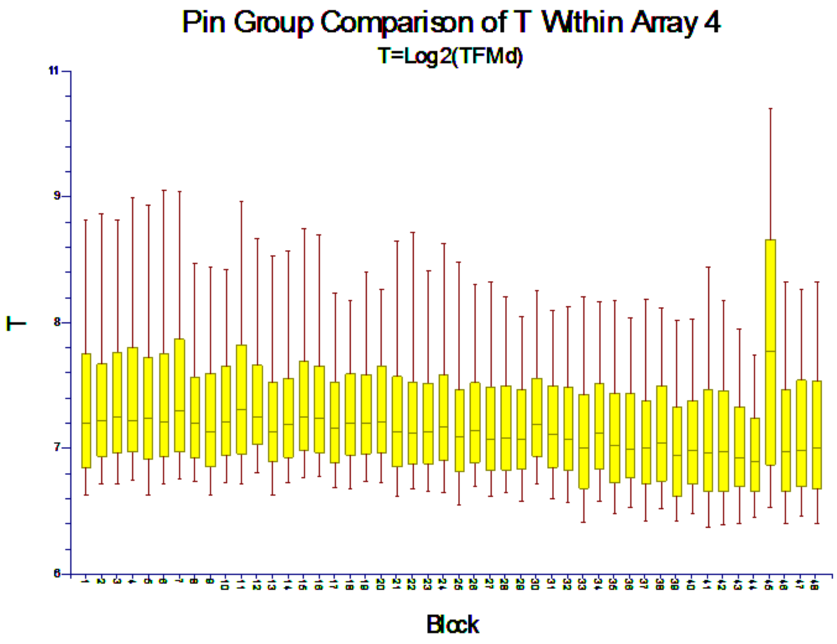

Whole Array Comparison of T

T=Log2(TFMd)

Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Pin (Print-tip) Group Box Plots

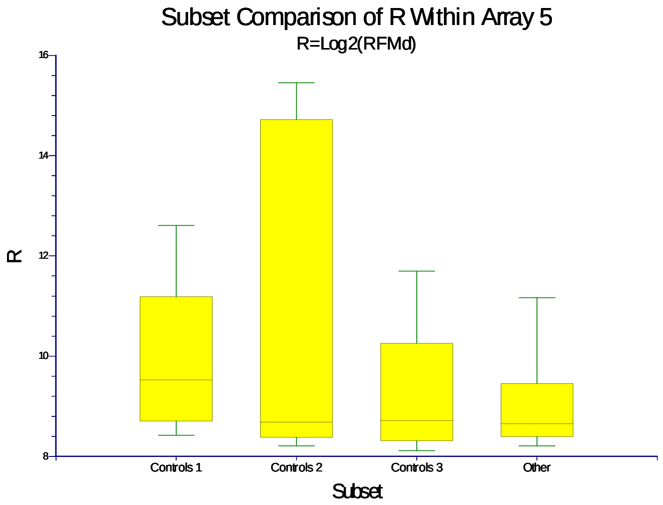
Spatial bias may also be seen by looking at box plots of the intensities of each of the pin (print-tip) groups. If one box plot is quite different from the others, it signals the print-tip (pin) used may be different from the others. The figure shows the box plots of Log2(median foreground intensities) of the Target sample. Pin (print-tip) group 45 may be suspect on this array.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Subset Box Plot

The intensity values of subsets can be compared to other subsets and non-subset spots using a subset side-by-side box plot. Often the subsets are various controls. The subset box plot below shows summaries of the Log2(reference foreground medians) for 3 controls and all other genes (spots).

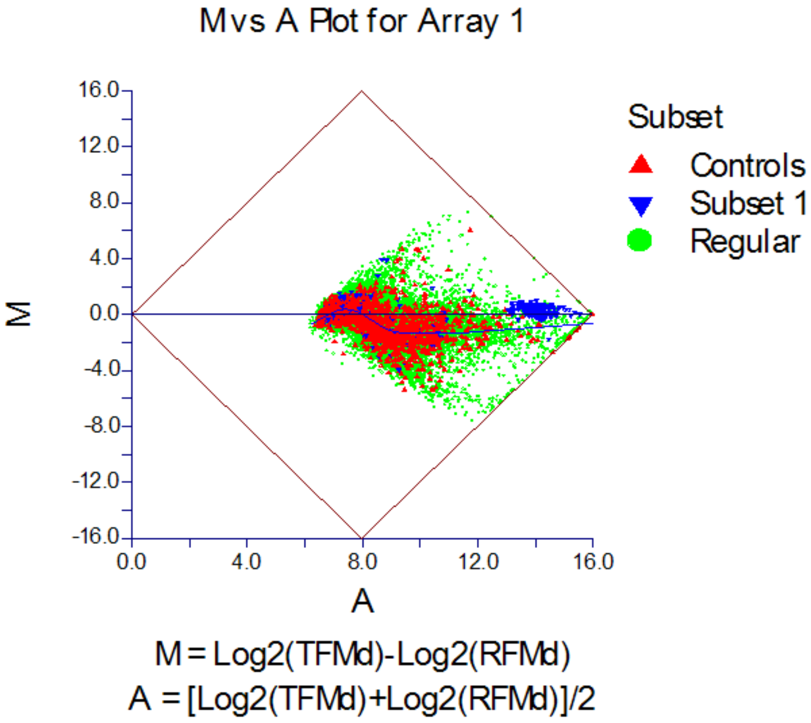


Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

M vs A Plot

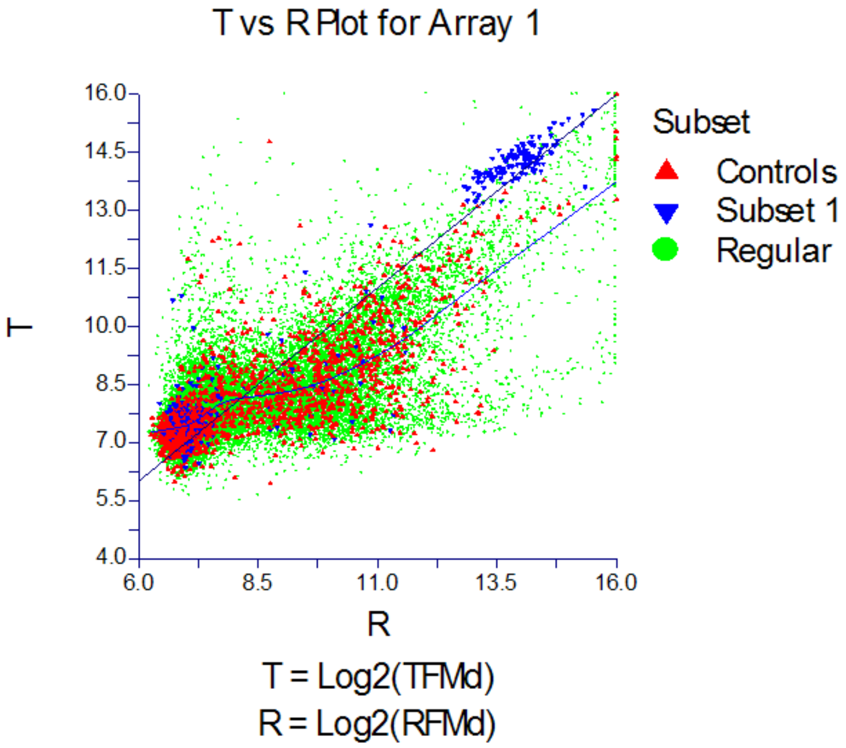
The M vs A plot is a scatterplot with a relative intensity measure (M) on the Y axis, and average intensity (A) on the X axis. This plot is useful for visualizing the relationship between dye-bias and intensity. Dye-bias occurs when one of the dyes produces a brighter image than the other dye for reasons other than differential expression. If there is no dye-bias, M values should be centered near zero across all values of A. If dye-bias exists, a banana shape is often the result. Plotting a loess line on the MA plot can be useful for viewing dye bias. A loess line is a specialized robust moving average that uses locally weighted polynomial regression. If the loess line is far from the zero line, there is evidence of dye bias. If the loess line is near the zero line, little or no dye-bias exists. An M vs A plot following loess normalization can be used to determine if the dye-bias problem has been properly corrected.

If control spots are shown in different colors, the M vs A plot is also useful for displaying how expression levels of control spots compare to those of the spots of interest. The diamond shows the physical limits of the M vs A coordinates when a 16-bit scanner is used. The overlaid loess line shows minor dye bias.



T vs R Plot

The relative intensity values of target and reference samples can be seen by plotting the target intensity measure on the Y axis, and the reference intensity measure on the X axis. This plot is similar to a rotation of the MA plot. This T vs R plot displays relative median foreground expression of target to reference samples. The upper physical limits of 16 can be seen here.



Numeric Summaries

Mean and standard deviation summaries for mean, median, or standard deviation of foreground and background regions of the spots are useful for comparing slides. Filter summaries across slides may also be good indicators of slides of questionable quality.

Step 8 – Spotted Array Individual Spot Quality and Filtering

Some useful indicators of individual spot quality include:

Standard deviation of Foreground or Background Pixels: This summary (of each channel) gives an idea of the uniformity of expression across the spot. Large standard deviations indicate non-uniformity. A large standard deviation can be determined by examining the whole slide numeric summary of standard deviations.

Flags: Some image processing software packages generate flags for spots which fail to meet some criteria. The criteria should be understood before using the flags as indicators of spot quality.

Saturation: Since each pixel has a limiting value of 65,535, pixels with such a value should be considered right-censored (unknown, but larger than 65,535). Spots with large proportions of saturated pixels may be termed saturated spots.

Weak Signal: Spots with low intensities are often near, or even below, the expression intensity of the background region. When expression intensities are near the background, it is difficult or impossible to estimate reliably the true expression level of that gene. It is only known that the gene is not highly or medium expressed. It is not known, however, whether the gene is expressed at very low levels, or not at all. Weak signal spots are essentially left-censored.

Weak Signal Considerations

Another aspect of foreground to background comparison should be mentioned here. Because the background is devoid of probe, it is not subject to nonspecific binding, as is the foreground, and thus may not accurately reflect the baseline it is intended to represent.

Many filters can be designed to remove values of questionable individual spot quality based on some cut-off value. However, it is much easier to flag a spot as one with questionable quality than it is to justify removal of the spot from future analysis. Spot filtering should not be treated lightly, as the following discussion illustrates.

An important decision that may have heavy bearing on the final analysis results is how the spots with low intensities will be treated. When the expression level of either channel is unknown, the relative expression of the two channels is also unknown. For example, suppose that for the Cyanine 3 (Cy3, green) channel at a given spot the estimated foreground is 100 and the estimated background is 150. The true expression intensity could be anywhere from 0 to 150 or more, but the techniques being used are not sensitive enough to distinguish a reliable estimate, due to background noise. Suppose further that the Cyanine 5 (Cy5, red) channel yields estimates of 600 for the foreground and 100 for the background intensity. Because 600 is well above 100, it can be assumed that the Cyanine 5 channel estimate is a reliable one. The difficulty in this situation lies in determining a good estimate of Cyanine 5 to Cyanine 3 relative intensity. The range of ratios is 600/150 = 4 to 600/0 = infinite. Even if logs are used, the ratio chosen will greatly affect the ensuing analysis, as will simply throwing the spot out, altogether.

Consider the following scenario in which the researcher throws out all weak-signal (or low relative intensity) spots. One thousand genes are to be compared based on 10 individuals each from a disease group and a non-disease group. It is of interest to determine which among the 1,000 genes is differentially expressed. Suppose there are four genes, which, unknown to the researcher, are highly expressed in the disease group, but not expressed in the non-disease group. The microarray experiment is run and the expression values for each of those four genes for all 10 individuals in the disease group are well above the background. However, the expression levels for those same four genes for the non-disease group are near background levels. Because the expression levels are near the background, the non-disease expression values for these four genes are all filtered from further analysis. A two-sample *t*-test is then attempted for each of the 1,000 genes. Unfortunately, the *t*-test cannot be run for the four genes of true differential expression, since for each of these four genes there are no expression values in the non-disease group. No evidence of differential expression is reported, and the four genes go undetected.

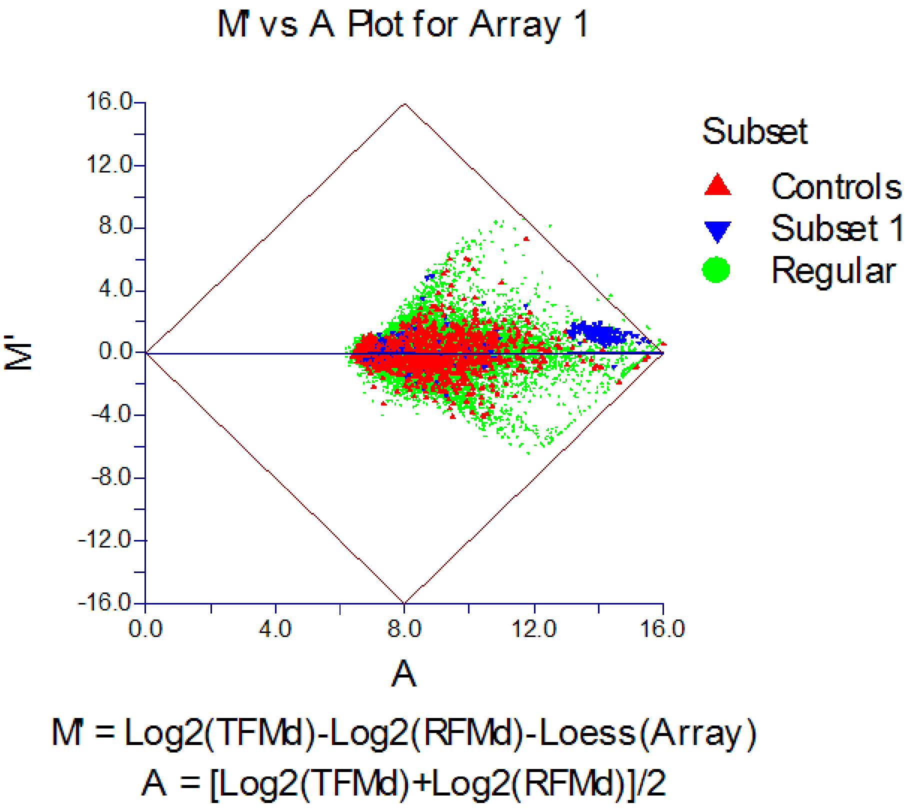
The preceding example reveals three aspects that should be considered when determining whether or not to filter a low intensity spot: (1) The foreground relative to the background for each dye individually; (2) The relative intensities of the two dyes at that spot; (3) The effect of removing that spot from the analysis, *across arrays*.

Step 9 – Array Normalization

The purpose of two-channel array (or pin/print-tip group) normalization is correcting for dye-bias. Dye-bias occurs when one of the dyes produces a brighter image than the other dye for reasons other than differential expression. Dye-bias is usually dependent on spot intensity and may be viewed using an MA plot with the loess line overlaid (see *MA Plot* under the heading *Two-Channel Whole Array Quality* above). To correct for intensity dependent dye bias at each *A* value, the loess value is subtracted from each *M* value at that location. The *M* values become normalized *M* values. This normalization causes the *M* values to be centered at zero along the horizontal axis (see the figure below) and the dye-bias has been removed. The loess normalization can be carried out for individual pin (print-tip) groups separately, or for the array as a whole.

M vs A Plot Following Array Normalization

Below is the M vs A plot of the same data used for the M vs A plot shown previously, but here following whole array loess normalization. The overlaid loess line follows the zero line, indicating the dye bias has been removed. Some points may go beyond the physical limits of the diamond following loess normalization.



GenePix® Files

GenePix Pro® is image analysis software that is often coupled with GenePix® scanners sold by Molecular Devices. Among its tools is a set of proprietary feature-finding algorithms, which are used to find and align each spot on the scanned array. Foreground and background regions are also determined. The foreground and background pixel intensities of each channel are summarized in a file with the suffix .gpr (gpr is the acronym for GenePix® Results). General spot description files may also be generated in a file with the suffix .gal (for GenePix Array List). GenePix® scanners and software are compatible with many commercially available slide arrays, including Combimatrix, Amersham, Illumina, Agilent, Schleicher & Scheull, and custom lab-printed arrays.

GPR Files

Using GenePix Pro®, a .gpr file is generated for each array that is scanned. Each .gpr file contains a header of several lines followed by the result columns. The number of lines in the result columns corresponds to the number of spots on the array. The first line of the result columns contains the column headings. The user of the GenePix Pro® software determines which columns are generated in the .gpr file. The choices of columns produced by the GenePix Pro® software have evolved with the versions of the software. For example, in GenePix Pro 6.0, up to 108 measurements can be generated for each spot. The following is a description of the columns which are commonly used in ***GESS***.

Block The block (or pin/print-tip group) of the spot.

Column The column of the spot, within the block.

Row The row of the spot, within the block.

Name The name of the spot, maximum of 40 characters.

ID The identifier of the spot, maximum of 40 characters.

X The horizontal coordinate of the center of the spot.

Y The vertical coordinate of the center of the spot.

F635 Median The median foreground pixel intensity for the red (635) channel.

F635 Mean The mean foreground pixel intensity for the red (635) channel.

F635 SD The standard deviation of the foreground pixel intensities for the red (635) channel.

B635 Median The median background pixel intensity for the red (635) channel.

B635 Mean The mean background pixel intensity for the red (635) channel.

B635 SD The standard deviation of the background pixel intensities for the red (635) channel.

F532 Median The median foreground pixel intensity for the green (532) channel.

F532 Mean The mean foreground pixel intensity for the green (532) channel.

F532 SD The standard deviation of the foreground pixel intensities for the green (532) channel.

B532 Median The median background pixel intensity for the green (532) channel.

B532 Mean The mean background pixel intensity for the green (532) channel.

B532 SD The standard deviation of the background pixel intensities for the green (532) channel.

% > B635 + 1 SD The percent of red (635) channel foreground pixel intensities which are greater than the red (635) channel background intensity plus one standard deviation.

% > B532 + 1 SD The percent of green (532) channel foreground pixel intensities which are greater than the green (532) channel background intensity plus one standard deviation.

F635 % Sat.	The percent of red (635) channel foreground pixel intensities which are the maximum intensity – 65,535.
F532 % Sat.	The percent of green (532) channel foreground pixel intensities which are the maximum intensity – 65,535.
Flags	Spot condition indicator: No flag (0), spot not found by automatic alignment (-50), spot is absent from GAL file (-75), spot is deemed bad by the user (-100), spot is deemed good by the user (100).

GAL Files

Because .gal files usually contain information that is general to all arrays in an experiment, there is usually only a single .gal file generated for the experiment. The format of a .gal file is similar to that of a .gpr file. Following a header of several lines are columns with information about each spot. While .gpr files give results about the reading of the array, the .gal file contains information about the purpose of the spot. Often, a column in a .gal file will indicate spots that are positive or negative controls, empty spots, or housekeeping genes, etc. A .gal file is also useful for detailed annotation of the genes of interest. The number of columns in a .gal file is usually much fewer than the number of columns in a .gpr file.

Two-Channel Designs

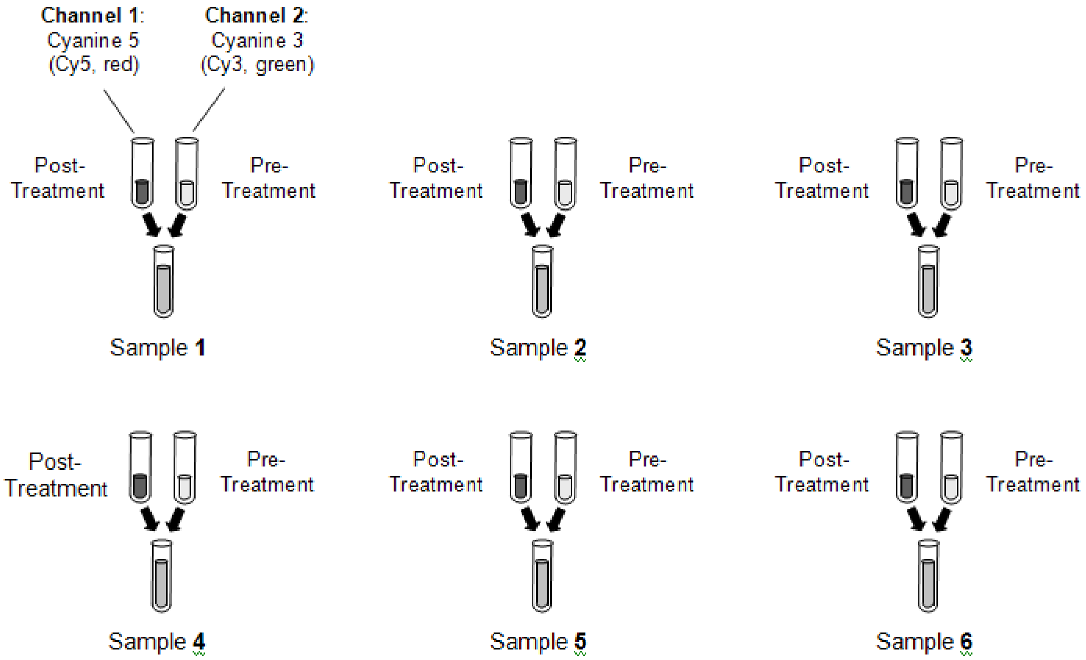
Two experimental designs may be used when using two channel microarrays: paired designs and reference designs.

Paired Design

The paired design is often used in two-channel experiments when the gene expression comparison to be made involves a natural pairing of experimental units.

As an example, suppose 6 cell samples are available for comparison. A portion of each of the 6 cell samples (before treatment) is reserved as a control. The same treatment is then given to each of the 6 remaining portions of the samples. It is of interest to determine the genes that are differentially expressed when the treatment is given. In this scenario, there is a natural before/after treatment pairing for each sample. The reserved control portions of each sample are labeled with Cyanine 3 (Cy3, green) dye, while the treatment portions are labeled with Cyanine 5 (Cy5, red) dye. From each sample, the labeled control and the labeled treatment portions are mixed and exposed to an array. The control and treatment portions compete to bind at each spot. The expression of treatment and control samples for each gene is measured with laser scanning. A pre-processing procedure is then used to obtain expression difference values for each gene. In this example, the result is 6 relative expression values (e.g., $\text{Log}_2(Post\ /\ Pre)$) for each gene represented on the arrays.

Paired Design, Six Arrays



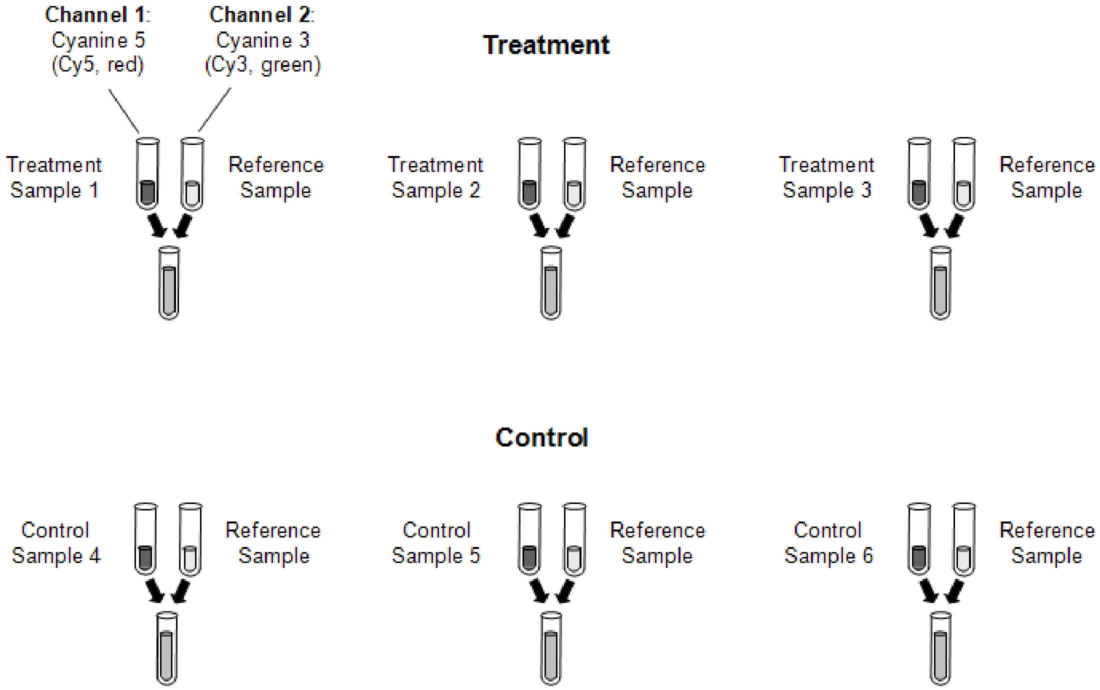
Reference Design

A two-sample reference design, or common reference design, employs an outside source of cDNA that is used as a reference for all samples in the experiment. Reference cDNA may be purchased separately or may be a combination of all cDNAs in the compared samples (The pros and cons of choice of reference cDNA is beyond the scope of this manual).

Suppose a treatment and control are to be compared. One group of experimental units serves as the control group. The other group of experimental units receives the treatment. Following treatment, cDNA is isolated for each of the experimental units. The cDNA for the treatment and control groups may be termed target cDNA. The target cDNA from both groups is labeled with Cyanine 5 (Cy5, red) dye. An outside source of cDNA, with (hopefully) most genes of interest expressed, is labeled with Cyanine 3 (Cy3, green) dye. This cDNA is the common reference, and is used as a baseline for all arrays of both groups. The intensity value for each gene of each array is the relative expression of the target cDNA to the reference cDNA at each spot (see data examples in the tables that follow).

The goal of the reference cDNA is to remove additional variation that may have been introduced in the experimental procedure. Array differences may be particularly pronounced when large periods of time pass between array hybridizations of a single experiment. Reference designs may also be employed in repeated measures/time-course designs.

Two-Sample Reference Design, Six Arrays

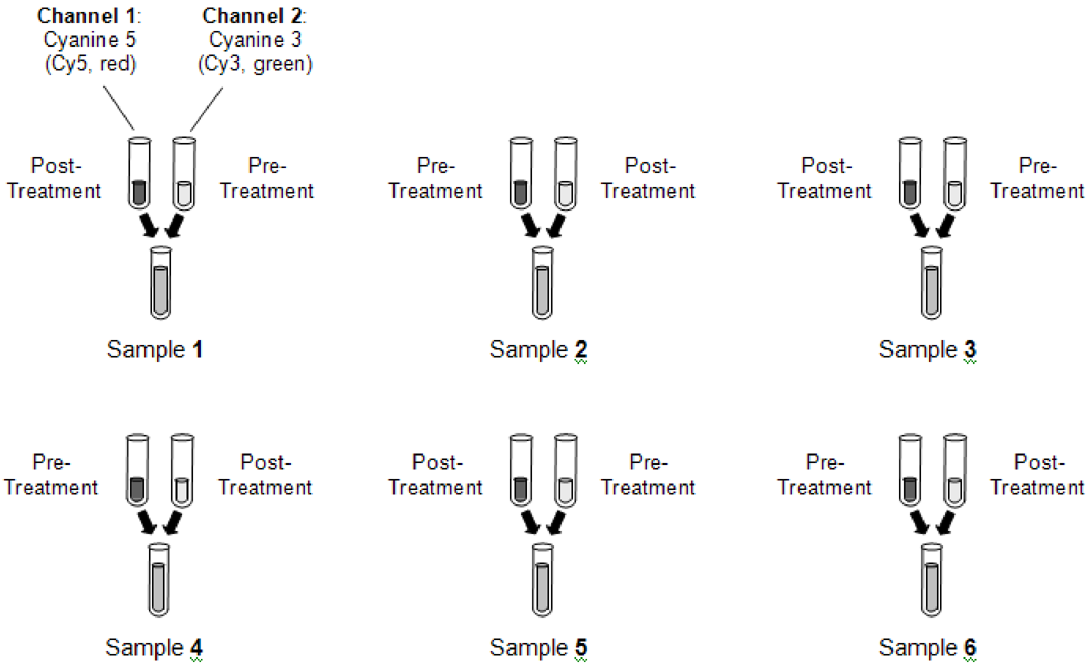


Dye Swap

Dye swap is a technique that may be employed in either paired or reference designs. The purpose is to remove systematic bias of the dye. To use this technique, the dye used is switched for a subset of the experimental units. For example, in the paired design example above, all 6 control portions are labeled with Cyanine 3 (Cy3, green) dye, while the 6 treatment portions are all labeled with Cyanine 5 (Cy5, red) dye. The dye swap technique could be employed by labeling half of the 6 controls with Cyanine 3 (Cy3, green) dye and the other 3 with Cyanine 5 (Cy5, red) dye. The 6 treatment portions would be labeled with the complement dye to that of the corresponding control portions. Careful record should be kept of which dyes are used on each array when performing an experiment with dye-swapping.

Dye-swap - Paired Design

Each of the 6 samples is divided into two portions. One portion serves as control. The other portion receives the treatment. Three of the treatment portions are labeled with Cyanine 5 dye. The corresponding control portions are labeled with Cyanine 3 dye. The other three treatment portions are labeled with Cyanine 3 dye. The corresponding control portions are labeled with Cyanine 5 dye. The samples are combined and then introduced onto a microarray slide. Six relative expression values are obtained for each probe: three are $\mathrm{Log}_2(Cy5 / Cy3)$, the other three are $\mathrm{Log}_2(Cy3 / Cy5)$.



Entering GPR Files

This section describes how file names are entered into the spreadsheet in preparation for preprocessing. Two variables (columns) are required to run GenePix pre-processing, and a third is required to obtain output files. A fourth variable, containing lists of bad pins, may also be used. Any name (entered by clicking on the Variable Info tab at the bottom of the spreadsheet) may be used for these variables.

GPR File Name Variable

The GPR file name variable is a column on the spreadsheet containing a list of paths and filenames of the .gpr files that are to be pre-processed. The files may be in different folders, but must all contain results for the same list of genes (spots). This variable is required to run the GenePix GPR File Pre-Processing Engine procedure.

Target Sample Dye Variable

When two-sample (two-channel) microarrays are used, one sample may be termed the target sample, while the other may be called the reference sample. It is important, particularly when the dye swap technique is used, but also in general, to keep track of whether the target sample is labeled with the Cyanine 5 (Cy5, red, 635) dye or the Cyanine 3 (Cy3, green, 532) dye. In ***GESS***, this is done with a target sample dye variable. A column is entered into the spreadsheet containing the dye (Cy5 or Cy3) of the target sample. Only the values Cy5 or Cy3 may be entered into this column. This variable is required to run the GenePix GPR File Pre-Processing Engine procedure.

Output File Names Variable

When the GenePix GPR File Pre-Processing Engine is run, a new set of files may be generated for use in statistical analyses. The path and name for these newly created files may be entered into the output file names variable or an empty column may be specified. The files may be in different folders. This variable is required to obtain output for statistical analysis.

Pin Groups to be Deleted Variable

When GenePix pre-processing indicates there are some pin (print-tip) groups of some arrays that are of such poor quality that they need be removed, this may be done by listing those pin groups in a column of the spreadsheet. Lists of the pin groups to be removed are entered into each cell, which corresponds to an output file, separated by spaces or commas.

For example, if it is desired that values from pin groups 12, 21, and 33 be filtered (removed) from the output file of Row 7, then *12 21 33* may be entered in the cell of Row 7 of the Pin Groups to be Deleted Variable column.

Procedure Options

This section describes the options available in this procedure. .

Variables Tab

These options specify the variables that will be used in the analysis.

GPR Input Files Specifications

These options are used specify the input .gpr files that are to be pre-processed.

GPR File Name Variable

Select the variable that contains the list of the .gpr array files of the experiment.

The names and pathways of the files should appear in a column below this variable name on the spreadsheet.

Target Sample Dye Variable

Select the variable that identifies whether the target sample is in the Cyanine 5 (Cy5, red) channel or the Cyanine 3 (Cy3, green) channel.

This required variable must contain either Cy5 or Cy3 in each cell.

The reference sample is automatically assumed to be in the other channel. That is, if the target sample is in the Cyanine 5 (Cy5, red) channel, the reference sample is assumed to be in the Cyanine 3 (Cy3, green) channel, and vice versa.

In some experiments, all target samples are in the Cyanine 5 (Cy5, red) channel. In other experiments, all reference samples are in the Cyanine 5 (Cy5, red) channel. In dye-swap experiments, the target and reference samples are mixed among channels.

In paired sample experiments, the reference and target samples may refer instead to before and after treatment.

Gene Name From

Specify which of the columns (NAME or ID) of the .gpr file will be used to identify the genes in the output file and ensuing statistical analyses.

Pixel Summary Statistic Settings

These options determine the summary value that will be produced in the output files.

Foreground Statistic

Specify whether the Median, Mean, or Standard Deviation of the pixel intensities is to be used.

For example, if Median is selected here, columns F635 Median and F532 Median of the .gpr file will be used.

NOTE: With F635 Median, F is for Foreground, 635 is the Cyanine 5 frequency, and Median indicates the median of all foreground pixel intensities is used.

Background Statistic

Specify whether the Median, Mean, or Standard Deviation of the pixel intensities is to be used.

For example, if Median is selected here, columns B635 Median and B532 Median of the .gpr file will be used.

NOTE: With B635 Median, B is for Background, 635 is the Cyanine 5 frequency, and Median indicates the median of all background pixel intensities is used.

Create Target and Reference Using

Specify how the target and reference samples will be summarized.

For example, if 'log2(Foreground - Background)' is specified here, the target samples will be summarized by taking the target foreground statistic, subtracting the target background statistic, followed by taking the logarithm, base2. A similar calculation will occur for the reference.

RECOMMENDATION: We recommend log2(Foreground).

Expression Measure that is Output

The formula selected here indicates the formula that will be used to summarize the expression at each spot of the array. These values are output into a file that can be used in the statistical analyses procedures.

Example: Suppose the target is in the Cy5 channel, 'Median' is selected under 'Foreground Statistic', 'log2(Foreground)' is selected under 'Create Target and Reference Using', and 'Target - Reference' is selected under 'Expression Measure that is Output'.

The output file will contain values using the formula log2(F635 Median) - log2(F532 Median), where F is for Foreground, 635 is the Cy5 (red) channel, and 532 is the Cy3 (green) channel.

RECOMMENDATION: We recommend Target - Reference - LOESS(Block).

- **Target**

  Only the target sample summary is output. The reference sample is ignored.

- **Reference**

  Only the reference sample summary is output. The target sample is ignored.

- **Target - Reference**
  The reference sample summary is subtracted from the background sample summary.

- **Target – Reference – LOESS(Array)**
  The loess value based on the entire array is subtracted.

- **Target – Reference – LOESS(Block)**
  The loess value based on the current block is subtracted.

GES Output Files Specifications

These options are used to determine the location and naming of the output .ges files.

Folder in which Output Files will be Stored

Enter the path and name of the folder in which the newly created .ges files will be stored. The path may be typed directly or the Browse button may be used to locate the desired folder. New files will be created only if a variable is entered under Output File Names Variable.

In the default path "%mydocs_NCSS%\DATA\GESS", the designator "%mydocs_NCSS%" represents the path to the ***GESS*** personal data folder (commonly *C:\...\[My] Documents\NCSS \NCSS 2007*).

Output File Names Variable

Select the variable in which the path and names of the newly created .ges files will be entered for future statistical analyses. The path and folder of these .ges files is entered under Folder in which Output Files will be Stored.

A new file (to be used for statistical analyses) will be created for each row when the procedure is run.

If this variable is left blank, no new .ges files will be created.

Append to File Names

A sequence of characters may be entered here to append to the name of the created file.

For example if the .gpr file has the name ‘Slide1_10hours.gpr’ and ‘log’ is entered here, the newly created .ges file will be ‘Slide1_10hours log.ges’.

If nothing is entered here, the file name will be the same as the name of the .gpr file, but ‘.gpr’ will be replaced with ‘.ges’.

Overwrite existing output (.ges) files with new output (.ges) files

Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file.

For example, if Slide1.ges has already been created and a new Slide1.ges file is to be written, the new file will be Slide1 (2).ges if the Overwrite box is not checked.

Reports Tab

The options on this panel control which reports and plots are generated.

Summary Reports

These options are used to determine the reports and report format that are output.

Specification Summary

Check this box to obtain a summary of the formula that is output to the output files, the subsets used, and the filters used.

Array Detail Summary

Check this box to obtain a row by row summary of names of files, numbers of filtered spots, array means and standard deviations, and deleted blocks (pin groups).

Mean Decimals

Specify the number of decimals used for means in the numeric portion of the Array Detail Summary. The number of decimal places is used for display only. It does not change the internal precision of the data.

SD Decimals

Specify the number of decimals used for standard deviations in the numeric portion of the Array Detail Summary. The number of decimal places is used for display only. It does not change the internal precision of the data.

Select Plots

The following options are used to determine which plots will be displayed.

Spatial Anomaly Plot

Check this box to indicate that you want this whole array spatial anomaly plot displayed for all arrays. The spatial anomaly plot gives a spatial view of the entire array for the corresponding measurement. Intensities are separated into four color groupings that reflect four percentile groups. The settings of the spatial anomaly plot are specified under the Spatial Plot tab.

Box Plot - Arrays

Check this box to indicate that you want to display side-by-side box plots comparing all arrays for this measurement. The settings of the box plot comparing array measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Box Plot - Pins

Check this box to indicate that you want to display side-by-side box plots comparing pin (print-tip) groups for this measurement for all arrays. The settings of the box plot comparing pin (print-tip) group measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Box Plot - Subsets

Check this box to indicate that you want to display side-by-side box plots comparing subsets (control groups) to the primary group of spots for this measurement for all arrays. The settings of the box plot comparing subsets (control groups) measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Scatter Plot – M vs A

Check this box to indicate that you want to display the M vs A plot for each array.

Sometimes called the Ratio-Intensity (R-I) plot, this plot is used to monitor dye bias. The Y axis specifies the value of M, the difference in the intensity summaries of the two samples, Target Summary - Reference Summary. M is a pneumonic for 'minus'. The X axis indicates the value of A, the average intensity summary of the two samples, (Target + Reference)/2. A is a pneumonic for 'add, or average, or abundance'. The loess line is a moving weighted regression average.

Scatter Plot – M’ vs A

Check this box to indicate that you want to display the M' vs A plot for each array.

This plot may be compared to the M vs A plot to see the effect of whole array loess subtraction on dye bias. The Y axis specifies the value of M', where M' = M - whole array loess value. The X axis indicates the value of A, which is the same as in the M vs A plot. The new loess line of the M' values is included.

Scatter Plot – M’’ vs A

Check this box to indicate that you want to display the M" vs A plot for each array.

This plot may be compared to the M vs A plot to see the effect of block loess subtraction on dye bias. The Y axis specifies the value of M", where M" = M - block loess value. The X axis indicates the value of A, which is the same as in the M vs A plot. The new loess line of the M" values is included.

Scatter Plot – T vs R

Check this box to indicate that you want to display the Target vs Reference plot for each array.

This plot is similar to the M vs A Plot. The Y axis shows the intensity summary of the Target sample. The X axis indicates the intensity summary of the Reference Sample.

Subsets 1 - 9 Tabs

The options on this panel control the names and lists of subsets.

Subset (1 – 9) Name

The name of the gene (spot) subset is entered here.

Plots comparing subsets can be obtained by checking the boxes next to 'Box Plot - Subsets' under the Reports tab. The name chosen here will appear on these plots.

EXAMPLE: To determine whether or not the microarray is functioning properly, it is common to introduce spike-in control DNA into the sample. Spike-in control DNA has a known relative intensity, e.g., 5-fold (target 5 times the reference sample), and corresponds to carefully chosen spots on the array. A list of Spike-in Controls could be given here. Values for the spike-in controls may be compared using box plots to negative controls, positive controls, blank spots, etc. to show that, in fact, the spike-in controls have higher relative intensity values. This would indicate a properly functioning microarray.

Genes in this Subset

Enter a list of genes (spots) that are to be in this subset. The genes (spots) may be entered directly, or the * character may be used to specify all genes with a particular beginning. The gene names or IDs entered in this list must be in the column specified in Spot Name From box on the Variables tab.

EXAMPLES:

Blank

spike1

spike3

spike5

spike* (all names beginning with spike)

AA44719

NM_00582

NM_04762

NM_27564

cntrl* (all names beginning with cntrl)

file(C:\Microarray\genelist.txt) (all names in the genelist.txt file)

var(OutputGenes) (all names in the spreadsheet variable with the variable name OutputGenes)

Output for Analysis

If this box is checked, the spots in this subset will be included in the output file for future statistical analyses. If this box is not checked, these spots will be removed from the output file.

(Plotting) Symbol

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the plotting symbol. This plotting symbol will be used in the selected array quality graphics.

Filters 1 Tab

The options on this panel control the weak signal filters that will be used.

Filter

Check this box to activate this filter. Spots meeting the criterion will be omitted from the output dataset, but will still be included in pre-processing graphical displays.

Filter Boundary

Specify the filter boundary value. When the spot value is below this boundary value, the output for this spot is omitted from the output dataset. All spots will still be included in pre-processing graphical displays.

Filters 2 Tab

The options on this panel control the saturation, standard deviation, pin group, and negative filters that will be used.

Saturation and SD Filter

Check this box to activate this filter. Spots meeting the criterion will be omitted from the output dataset, but will still be included in pre-processing graphical displays.

Saturation and SD Filter Boundary

Specify the filter boundary value. When the spot value is above this boundary value, the output for this spot is omitted from the output dataset. All spots will still be included in pre-processing graphical displays.

Pin Groups to be Deleted Variable

Specify the name of the variable that contains a list of pin (print-tip) groups to be deleted from the output dataset. This variable should have been created on the spreadsheet. Pin (print-tip) groups are excluded within arrays according to this variable, not across arrays. Pin (print-tip) groups should be separated by spaces as entries in the cells of the column under this variable name.

This variable is usually created after an initial run to identify problematic pin (print-tip) groups.

Delete spot if it has a negative flag

Check this box to activate this filter. If this box is checked, spots for which a negative flag was generated when creating the .gpr file will be omitted from the output dataset, but will still be included in pre-processing graphical displays.

The following are the possible negative value flags and their meanings:

Flag Value (Meaning)

-50 (Spot not found by automatic alignment)

-75 (Spot is absent from .gal file)

-100 (Spot is deemed bad by the user)

Spatial Plot Tab

The options on this panel control the features of the spatial anomaly plot.

Heat Map Settings

These settings are used to control the appearance of the heat map and its legend.

Heat Map Colors and Scale

Click on the heat map color bar or the button to the right to change the colors and/or scale of the heat map.

Label

Enter text here for the legend label.

Number of Values

This is the number of reference values printed along the right side of the heat map legend.

Show Legend

Specify whether to show the legend.

Value Format

This option specifies the characteristics of the reference numbers shown next to the heat map legend.

It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot.

It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Plot Settings

These options are used to specify the appearance of the heat map surroundings.

Plot Style file

A plot style file sets all plot options that are not set directly by this procedure.

Interior Color

Specify the interior color of the spatial anomaly plot.

Background Color

Specify the background color of the spatial anomaly plot.

Plotting Symbol Width and Height

This is the width and height in thousandths of an inch of the rectangle that is plotted for each gene.

Recommended:

Width = 120

Height = 150

Top and Bottom Titles

Plot Titles

Enter text here for the designated title or label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by the long version of the selected intensity summary.

{Y} is replaced by the short version of the selected intensity summary.

{Z} is replaced by the appropriate array number.

Box Plot Tab

The options on this panel control attributes of the box plots.

Vertical and Horizontal Axes

These options control the vertical and horizontal axes attributes.

Axis Labels

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by the appropriate values of the corresponding X axis grouping variable.

{Y} is replaced by the short version of the Y axis intensity summary.

Ref. Number Format

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Major Ticks

Specify the number of large tickmarks and optional grid lines along this axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along this axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along this axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the axis. Thus, checking the Y Grid Lines will actually cause horizontal grid lines to appear.

Box Plot Settings

These options are used to specify the appearance of the box plots.

Box Plot Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created in the Box Plots procedure.

Box Percent Space

When the Box Width (or Bar Width) option is set to Percent Space in the Box Plot Style File selected, this value specifies the percent of the length of the axis that is empty space instead of

bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**

  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**

  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Interior Color

Specify the interior color of the plot.

Background Color

Specify the background color of the plot.

Titles

Enter text for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{G} is replaced by the long version of the Y axis intensity summary.

{Y} is replaced by the short version of the Y axis intensity summary.

{Z} is replace by the appropriate array number.

Box Plot Colors

The options are used to specify the colors of the box plots.

Fill Color

The color used to fill this object. Click to change.

Outline (Border) Color

The color used to outline the object. Click to change.

Line Color

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Scatter Plot Tab

The options on this panel control the main features of the M vs A and T vs R plots.

Vertical and Horizontal Axes

These options control the vertical and horizontal axes attributes.

Ref. Number Format

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Major Ticks

Specify the number of large tickmarks and optional grid lines along this axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along this axis.

Grid Lines

Check this option to display grid lines at the major tickmarks along this axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the axis. Thus, checking the Horizontal Axis Grid Lines will actually cause vertical grid lines to appear.

Scatter Plot Settings

These options are used to specify the appearance of the scatter plots.

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. Scatter plot style files are created in the Scatter Plots procedure.

Symbol

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the points plotted on the scatter plot.

Diamond

Check this box to display a diamond showing the physical boundaries of M on the M vs A, M' vs A, and M" vs A scatter plots.

Zero

Check this box to display a horizontal line at 0 on the M vs A, M' vs A, and M" vs A scatter plots.

45 Degree

Check this box to display a 45 degree line on the T vs R scatter plot only. The 45 degree line shows where T and R are equivalent.

Interior Color

Specify the interior color of the plot.

Background Color

Specify the background color of the plot.

Plot Titles

Enter text here for the designated title or label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{M} is replaced by the long version of the X axis intensity summary.

{S} is replaced by the long version of the Y axis intensity summary.

{X} is replaced by the short version of the X axis intensity summary.

{Y} is replaced by the short version of the Y axis intensity summary.

{Z} is replaced by the appropriate array number.

Loess Options

These options are used to determine whether a loess line is included, its appearance, and the details of how it is computed.

Include Loess Curve

Check this option to display a Loess smooth line.

The locally-weighted, robust regression (loess) smooth is a popular, computer-intensive technique that usually provides a reasonable smoothing of your data without being overly sensitive to outliers. A reasonable smooth is one that travels more or less through the middle of the data. The degree of smoothing is controlled by the Loess % N option.

Loess Order

The order of the polynomial fit in the Loess procedure. Select '1' for a linear fit or '2' for a quadratic fit.

RECOMMENDED: 2 - Quadratic

Loess % N

The percent of the dataset to be used at each Loess calculation.

RECOMMENDED: 40

RANGE: 1 to 99

Number of Points

Specify the number of points at which the Loess line is evaluated. This affects the granularity of the lines. More points imply smoother lines. The number of points selected here may considerably affect the run time.

RANGE: 20 to 2000.

Min-Max Tab

The options on this panel control the minimum and maximum values for the axes of the box plots and scatter plots.

Axis Minimum

Specify the value to be displayed as the minimum on this axis. Data values less than this amount will be ignored. If this value is left blank, the minimum will be determined from the data.

Axis Maximum

Specify the value to be displayed as the maximum on this axis. Data values greater than this amount will be ignored. If this value is left blank, the maximum will be determined from the data.

Labels Tab

The options on this panel control the labels used for scatter plots, spatial anomaly plots, and box plots.

Short Label

Enter here the text that is to be used for the short labels of box plots, spatial anomaly plots, and scatter plots.

Long Label

Enter here the text that is to be used for the long labels of box plots, spatial anomaly plots, and scatter plots. The default is based on the entries under Pixel Summary Statistic Settings of the Variables Tab.

Setup Tab

This panel is used to specify the column headings that are to be read from the .gpr files.

Column Names

A GenePix compatible (.gpr) file is made up of many columns. Each column has a corresponding heading name. This program uses the heading name specified here to search for the appropriate column to read from the GenePix file. The option to change this name is provided so you can change the column heading name to be read, should the need arise.

If you change this heading name to one that does not exist in the file, that column will not be read.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Pre-Processing GenePix GPR Files

This section presents an example of how to pre-process six .gpr files, without involving subsets or filters. The spreadsheet data used are recorded in the GP_Ex1 dataset.

To run this example, take the following steps or load the **Example 1** template from the GenePix GPR File Pre-Processing Engine Template tab.

1 Open the GP_Ex1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **GP_Ex1.S0**.
- Click **Open**.

2 Open the GenePix Pre-Processing Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **GenePix GPR Files**. The GenePix GPR File Pre-Processing Engine procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the GenePix GPR File Pre-Processing Engine window, select the **Variables tab**.
- Set the **GPR File Name Variable** to **InputFile**.
- Set the **Target Sample Dye Variable** to **Target**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **OutputFile**.
- Leave **Append to File Names** blank.

4 Specify the pixel summary statistic settings.

- Continuing on the Variables tab, set the **Foreground Statistic** to **Median**.
- Set **Create Target and Reference Using** to **log2(Foreground)**.
- Set **Expression Measure that is Output** to **Target – Reference – LOESS(Block)**.

5 Specify the reports.

- Select the **Reports tab**.
- Check the boxes next to **Specification Summary** and **Array Detail Summary**.
- Check the **Cy5** and **Cy3** boxes next to S**patial Anomaly Plot**, **Box Plot – Arrays**, and **Box Plot – Pins**.
- Check the **M** and **M’’** boxes to the right of **Scatter Plot – vs A**.

6 Specify the spatial anomaly plot settings.

- Select the **Spatial Plot tab**.
- Change the **plotting symbol width** to **70**, and the **height** to **140**.

7 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Expression Formula Output for Analysis

```
Expression Formula Output for Analysis

Log2(TFMd)-Log2(RFMd)-Loess(Block)

where

Log2: Logarithm Base 2
TF: Target Sample, Foreground Region of Spot
RF: Reference Sample, Foreground Region of Spot
Md: Median Pixel Intensity
Loess(Block): Loess Values Calculated Within Each Block
```

This report displays the formula that is used when the output files are created. In this case, the formula may be read as ‘log base 2 of the target foreground median minus log base 2 of the reference foreground median minus the within block loess value.

Value for Spot Was Deleted (Filtered) If

```
Value for Spot Was Deleted (Filtered) If

No Filters Selected
```

This report shows that no filters were selected.

Subset Summary

Subset	Subset Values Output for Analysis?

No subsets were used.

This report shows that no subsets were created.

Input File Summary

Input File Summary

Row	Input File
1	…\Data\GESS\GP\Slide 1.gpr
2	…\Data\GESS\GP\Slide 2.gpr
3	…\Data\GESS\GP\Slide 3.gpr
4	…\Data\GESS\GP\Slide 4.gpr
5	…\Data\GESS\GP\Slide 5.gpr
6	…\Data\GESS\GP\Slide 6.gpr

This report shows a list of the input file paths.

Output File Summary

Output File Summary

Row	Output File
1	…\data\gess\Slide 1.ges
2	…\data\gess\Slide 2.ges
3	…\data\gess\Slide 3.ges
4	…\data\gess\Slide 4.ges
5	…\data\gess\Slide 5.ges
6	…\data\gess\Slide 6.ges

This report shows a list of the output file paths. These are the names of the files that will be used as input for statistical analyses.

Numeric Array Summary - Foreground

Numeric Array Summary - Foreground

			Standard		Standard
		Mean	Deviation	Mean	Deviation
		of Target	of Target	of Reference	of Reference
		Foreground	Foreground	Foreground	Foreground
Row	Input File Name	Medians	Medians	Medians	Medians
1	Slide 1.gpr	868.4	3471.4	1657.4	6308.9
2	Slide 2.gpr	1403.9	5359.7	2027.5	7169.7
3	Slide 3.gpr	1061.9	4828.0	1987.8	7039.0
4	Slide 4.gpr	694.0	3175.7	1812.6	6582.1
5	Slide 5.gpr	1239.9	5147.0	2831.4	8577.5
6	Slide 6.gpr	1814.6	7468.7	5290.9	12102.2

This report shows the whole array foreground region means and standard deviations for target and reference samples.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Mean of Target Foreground Medians

For the target sample, this is the average of all median pixel intensities of the foreground regions of the entire array.

Standard Deviation of Target Foreground Medians

For the target sample, this is the standard deviation of all median pixel intensities of the foreground regions of the entire array.

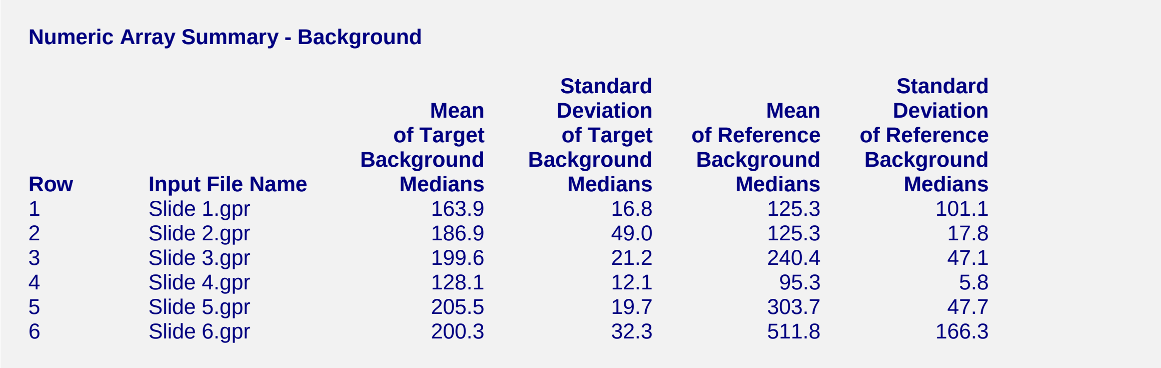
Mean of Reference Foreground Medians

For the reference sample, this is the average of all median pixel intensities of the foreground regions of the entire array.

Standard Deviation of Reference Foreground Medians

For the reference sample, this is the standard deviation of all median pixel intensities of the foreground regions of the entire array.

Numeric Array Summary - Background



Numeric Array Summary - Background

Row	Input File Name	Mean of Target Background Medians	Standard Deviation of Target Background Medians	Mean of Reference Background Medians	Standard Deviation of Reference Background Medians
1	Slide 1.gpr	163.9	16.8	125.3	101.1
2	Slide 2.gpr	186.9	49.0	125.3	17.8
3	Slide 3.gpr	199.6	21.2	240.4	47.1
4	Slide 4.gpr	128.1	12.1	95.3	5.8
5	Slide 5.gpr	205.5	19.7	303.7	47.7
6	Slide 6.gpr	200.3	32.3	511.8	166.3

This report shows the whole array background region means and standard deviations for target and reference samples.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Mean of Target Background Medians

For the target sample, this is the average of all median pixel intensities of the background regions of the entire array.

Standard Deviation of Target Background Medians

For the target sample, this is the standard deviation of all median pixel intensities of the background regions of the entire array.

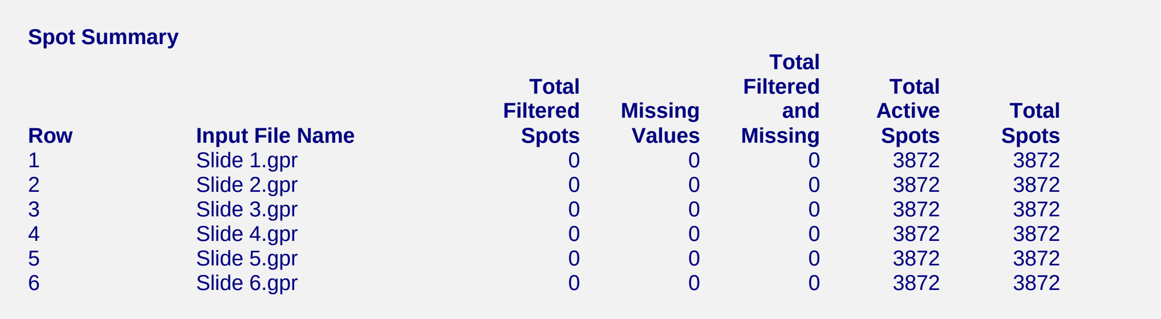
Mean of Reference Background Medians

For the reference sample, this is the average of all median pixel intensities of the background regions of the entire array.

Standard Deviation of Reference Background Medians

For the reference sample, this is the standard deviation of all median pixel intensities of the background regions of the entire array.

Spot Summary



Spot Summary

				Total		
		Total		Filtered	Total	
		Filtered	Missing	and	Active	Total
Row	Input File Name	Spots	Values	Missing	Spots	Spots
1	Slide 1.gpr	0	0	0	3872	3872
2	Slide 2.gpr	0	0	0	3872	3872
3	Slide 3.gpr	0	0	0	3872	3872
4	Slide 4.gpr	0	0	0	3872	3872
5	Slide 5.gpr	0	0	0	3872	3872
6	Slide 6.gpr	0	0	0	3872	3872

This report shows a summary of filtered spots, missing values, active and total spots.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Total Filtered Spots

This is number of spots that were filtered. The specifics of the filter are found in the next summary.

Missing Values

This is number of missing values among all spots.

Total Filtered and Missing

This is the sum of the Total Filtered Spots and the Missing Values.

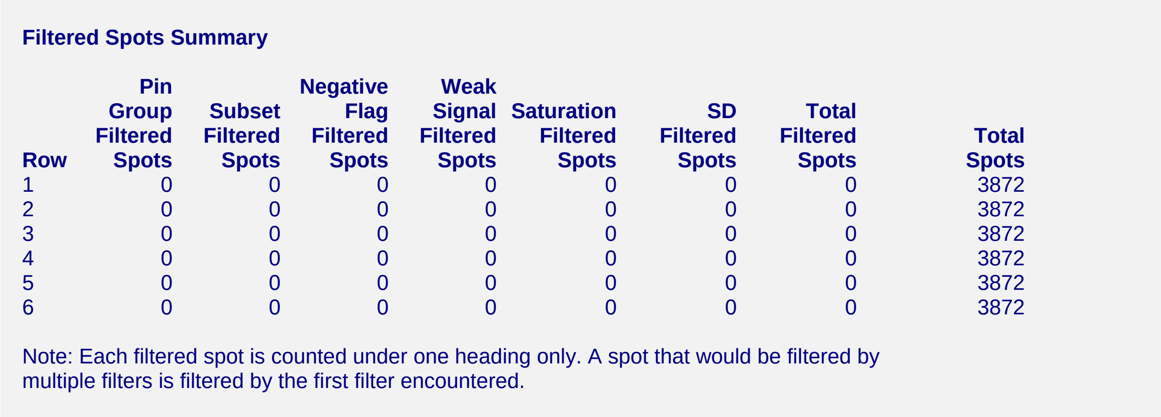
Total Active Spots

This is the number of spots that are not filtered, nor missing.

Total Spots

This is the total number of spots on the array.

Filtered Spots Summary



Filtered Spots Summary

Row	Pin Group Filtered Spots	Subset Filtered Spots	Negative Flag Filtered Spots	Weak Signal Filtered Spots	Saturation Filtered Spots	SD Filtered Spots	Total Filtered Spots	Total Spots
1	0	0	0	0	0	0	0	3872
2	0	0	0	0	0	0	0	3872
3	0	0	0	0	0	0	0	3872
4	0	0	0	0	0	0	0	3872
5	0	0	0	0	0	0	0	3872
6	0	0	0	0	0	0	0	3872

Note: Each filtered spot is counted under one heading only. A spot that would be filtered by multiple filters is filtered by the first filter encountered.

This report shows a detailed summary of all filtered spots.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Pin Group Filtered Spots

This is the total number of spots that were filtered because they were members of a deleted pin group.

Subset Filtered Spots

This is the total number of spots that were filtered because they were members of a deleted subset.

Negative Flag Filtered Spots

This is the total number of spots that were filtered because they had an associated negative flag.

Weak Signal Filtered Spots

This is the total number of spots that were filtered based on one or more of the twelve weak signal filters.

Saturation Filtered Spots

This is the total number of spots that were filtered based on one or more of the two saturation filters.

SD Filtered Spots

This is the total number of spots that were filtered based on one or more of the four standard deviation filters.

Total Filtered Spots

This is number of spots that were filtered.

Total Spots

This is the total number of spots on the array.

Deleted Pin Groups Summary

Filtered Spots Summary

Row	Input File Name	Deleted Pin Groups
1	Slide 1.gpr	
2	Slide 2.gpr	
3	Slide 3.gpr	
4	Slide 4.gpr	
5	Slide 5.gpr	
6	Slide 6.gpr	

This report shows a summary of pin groups that were filtered within each array.

Row

This is the row of the array in the spreadsheet.

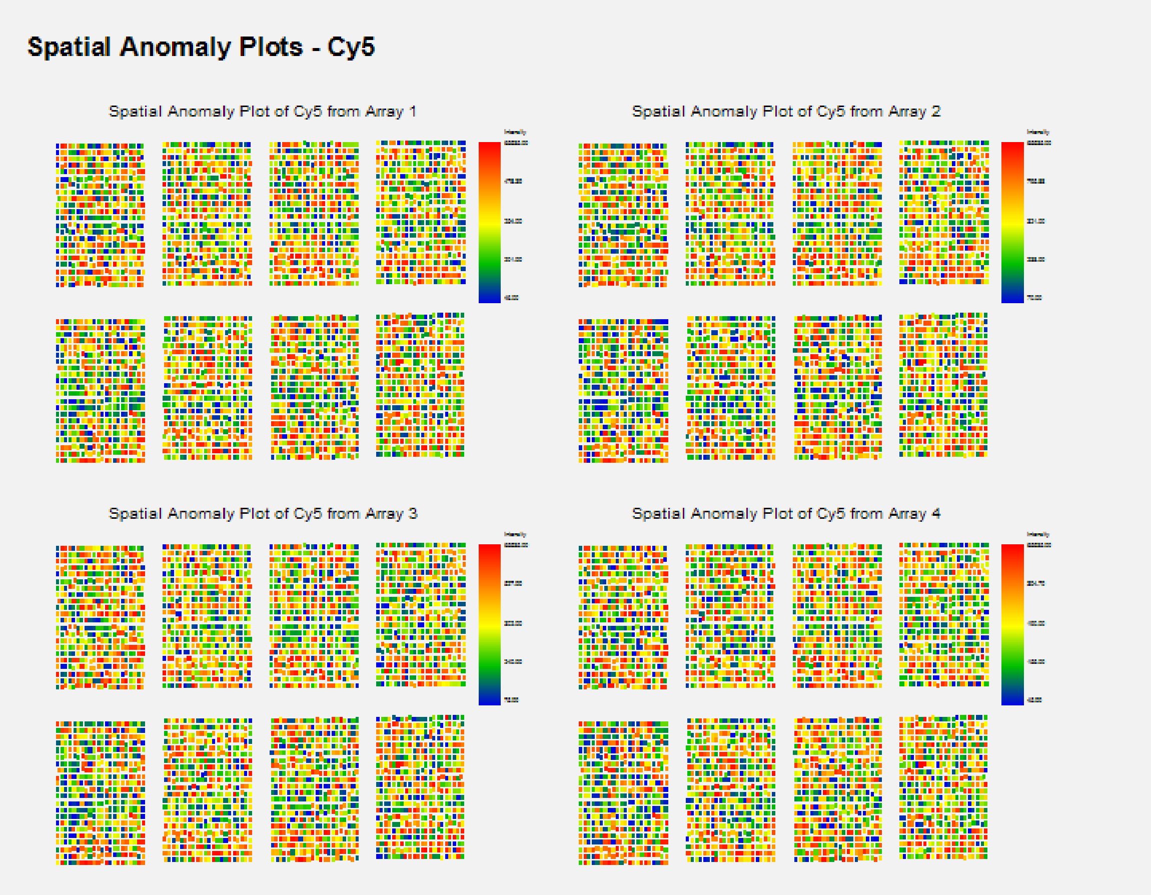
Input File Name

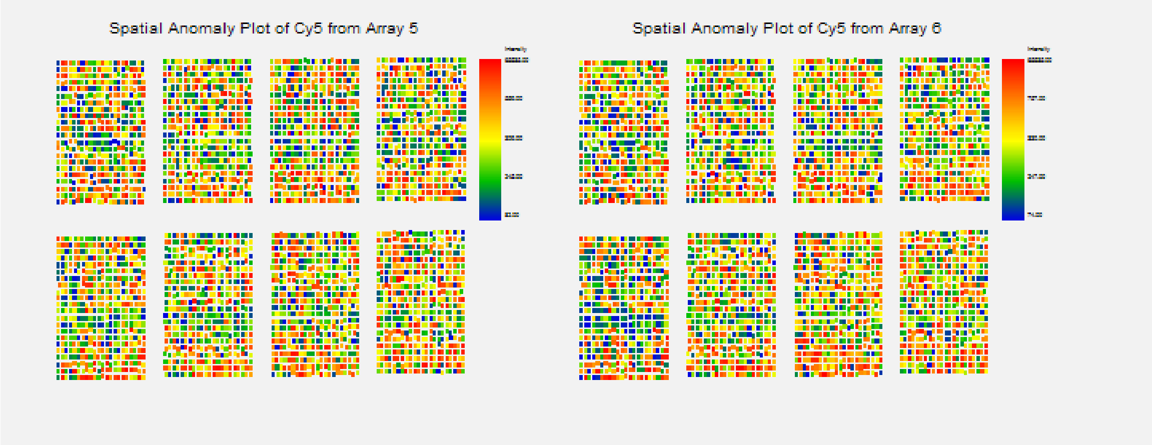
This is the name of the file without the path.

Deleted Pin Groups

These are the pin groups for which the all spots are deleted.

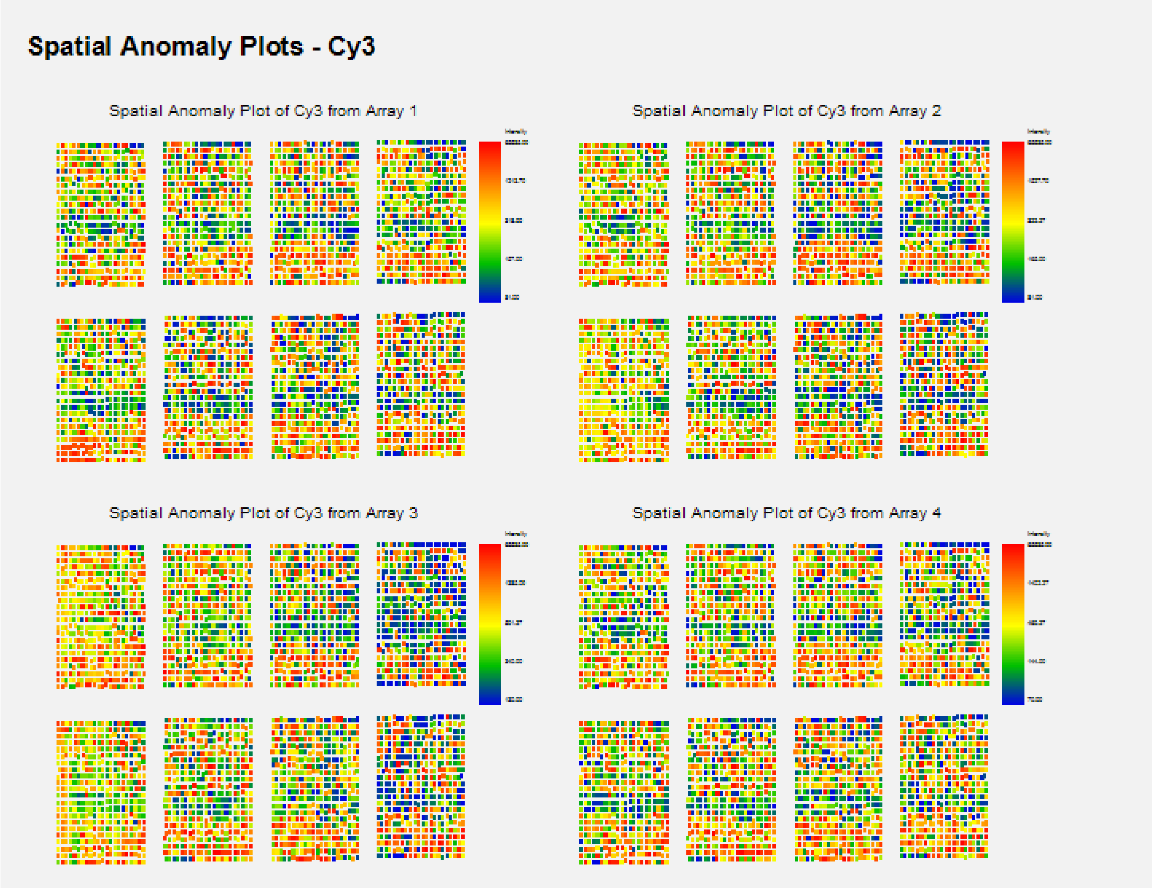
Spatial Anomaly Plots – Cy5

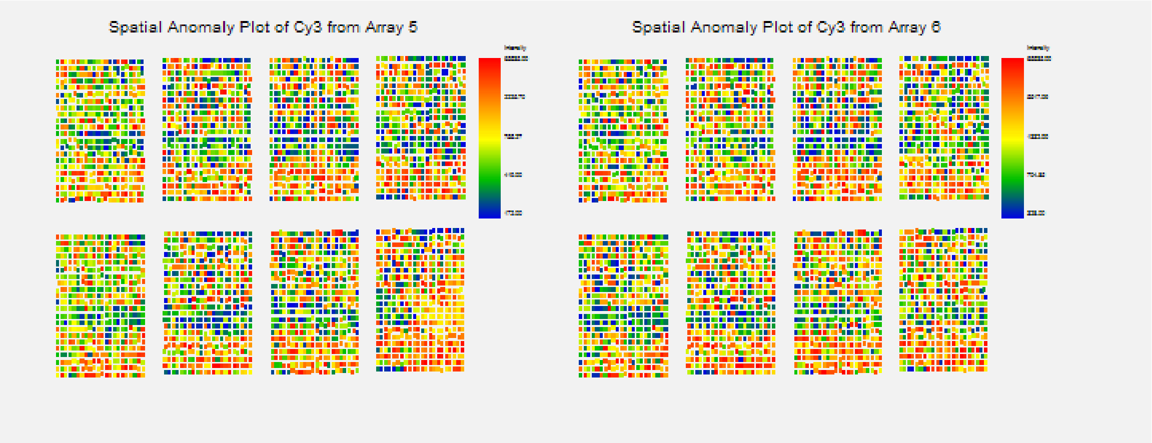




This report shows a spatial representation of the Cyanine 5 (Cy5, red) median foreground intensities. There appears to be a striped pattern on most of the arrays.

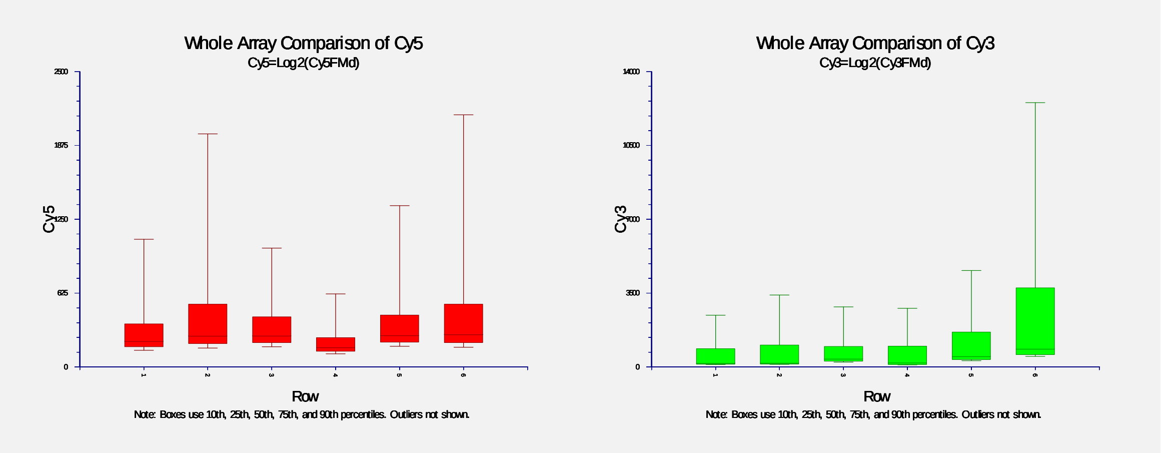
Spatial Anomaly Plots – Cy3





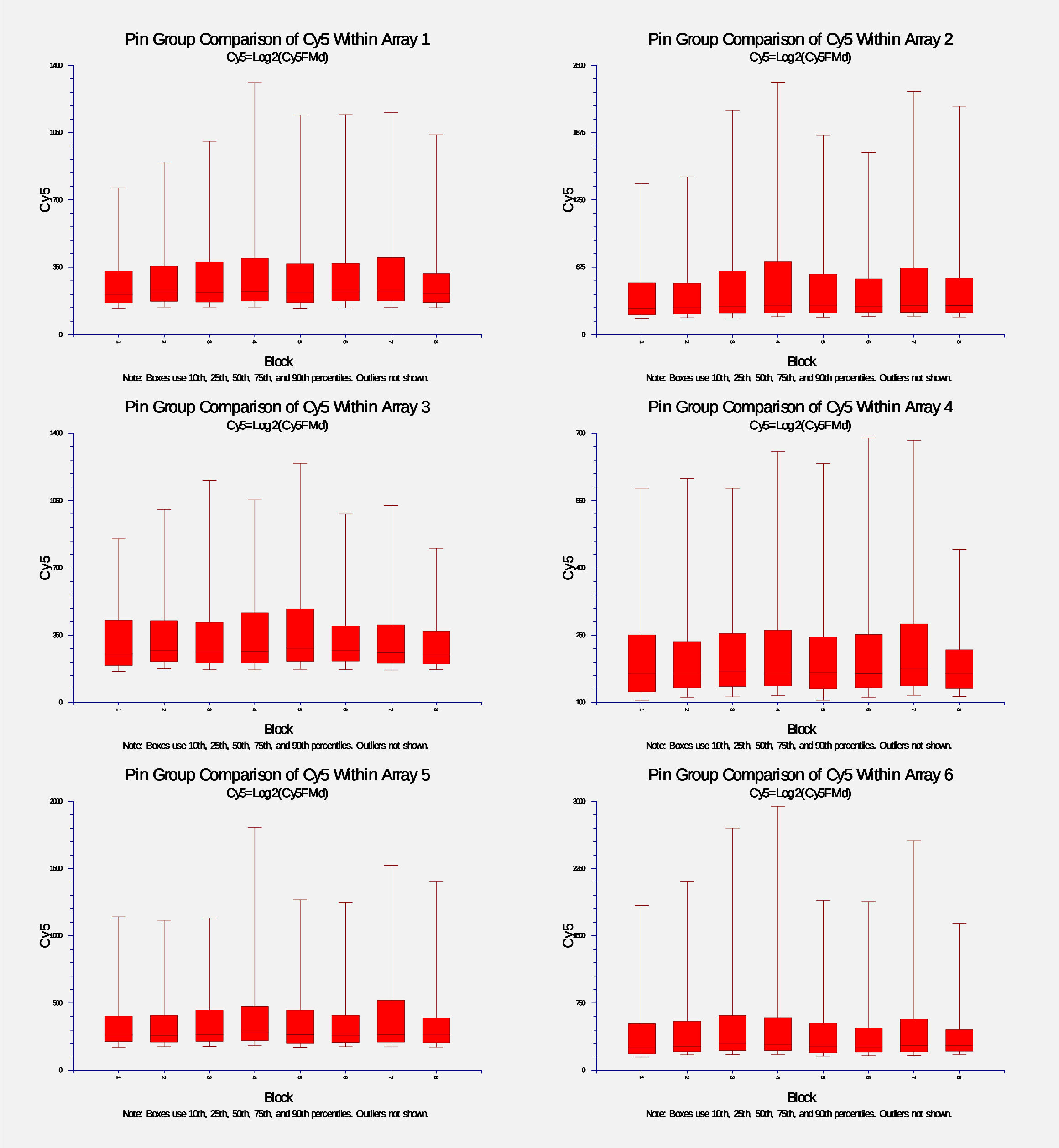
This report shows a spatial representation of the Cyanine 3 (Cy3, green) median foreground intensities. A similar striped pattern appears with Cy3. There are also some regions of unusual brightness on some of the arrays, such as in the 8th pin group of Array 5.

Array Comparison Section



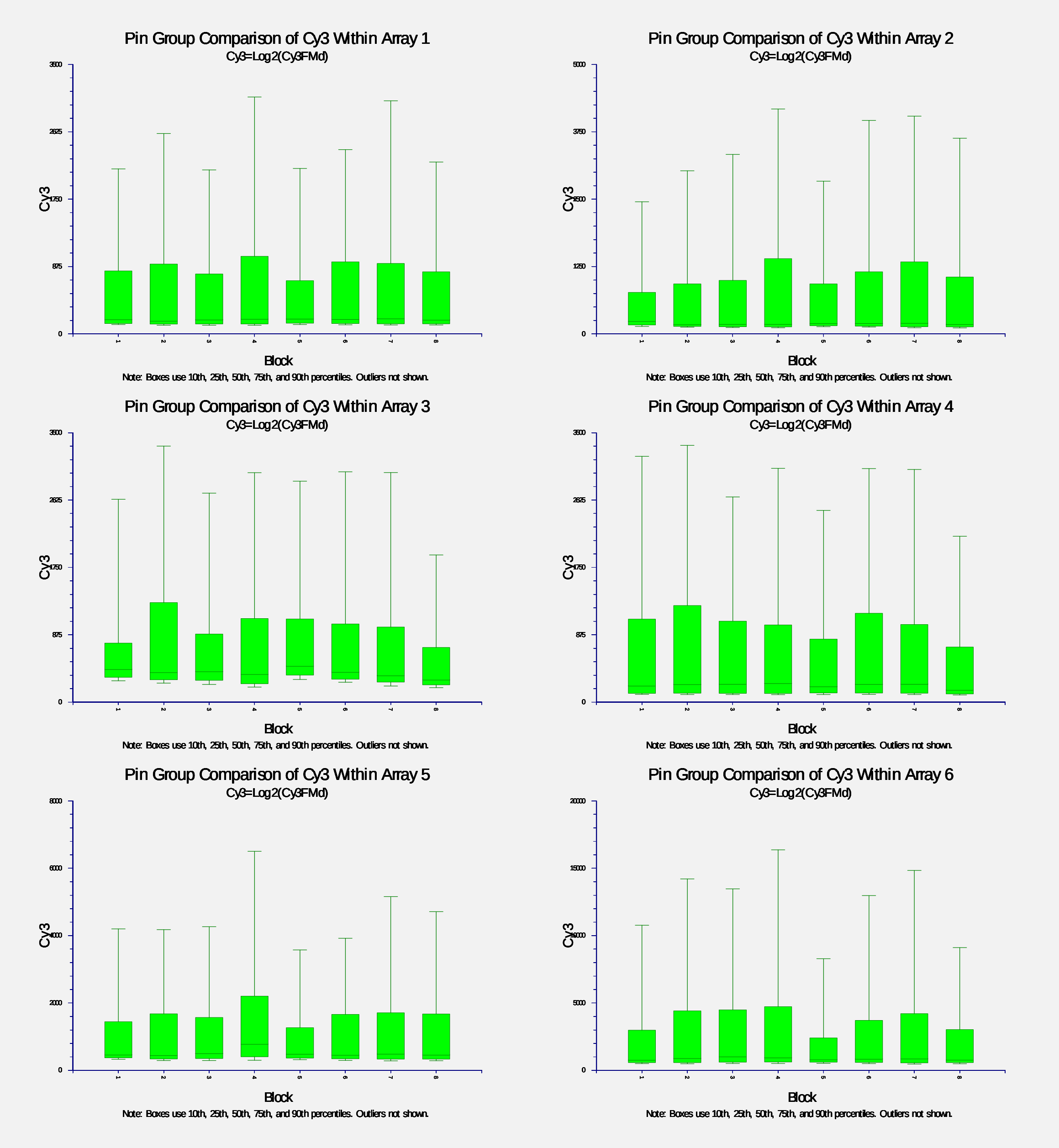
These plots allow comparison of Log2(Foreground Median) of the 6 slides for both Cyanine 5 (Cy5, red) and Cyanine 3 (Cy3, green) channels. The expression of the Cy3 channel is unusually high in Slide 6.

Pin Group Comparison Section – Cy5



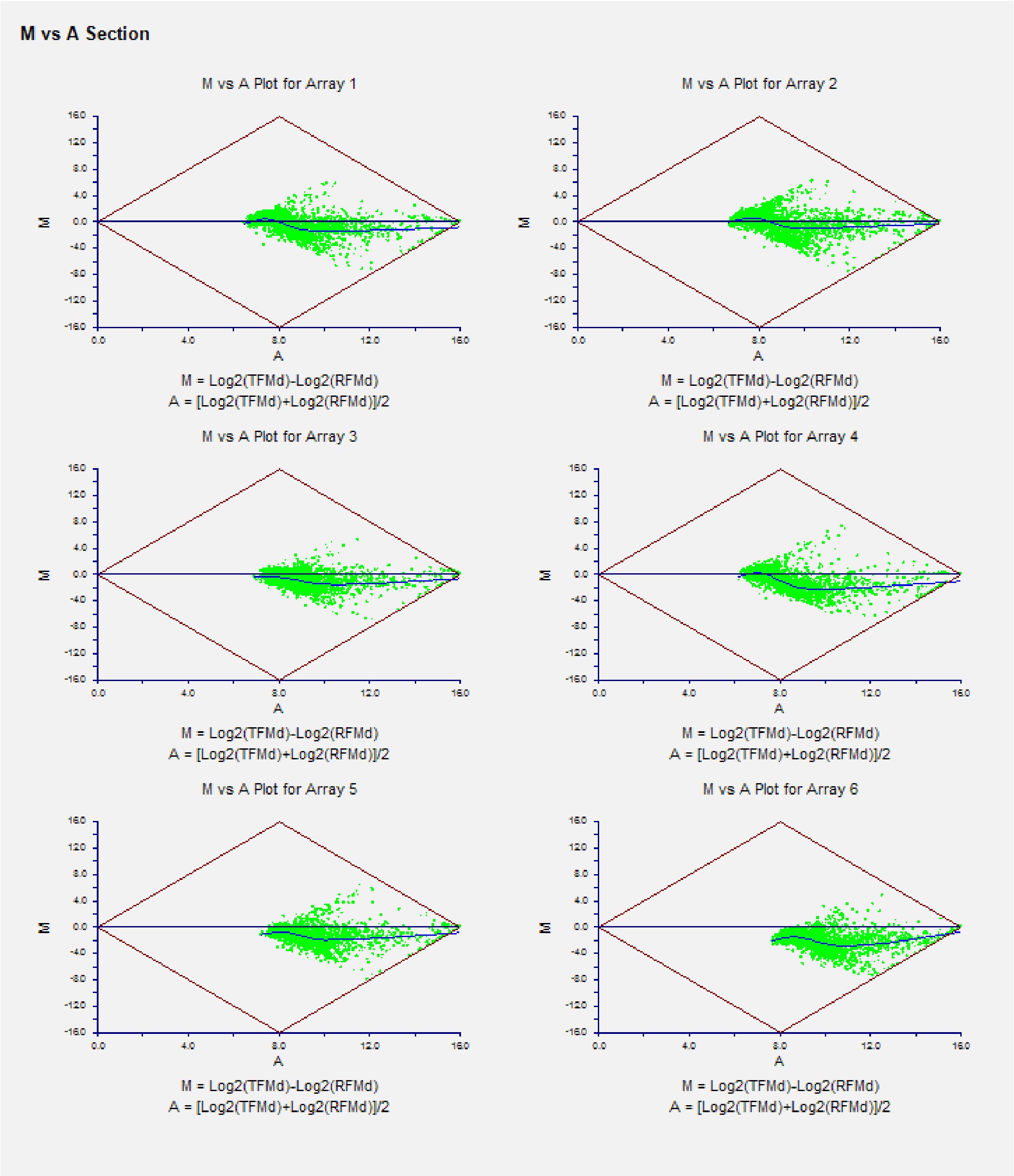
These plots allow comparison of Log2(Foreground Median) across all pin groups within each slide for the Cyanine 5 (Cy5, red) channel.

Pin Group Comparison Section – Cy3



These plots allow comparison of Log2(Foreground Median) across all pin groups within each slide for the Cyanine 3 (Cy3, green) channel.

M vs A Section



M vs A Section

M vs A Plot for Array 1

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 2

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 3

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 4

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 5

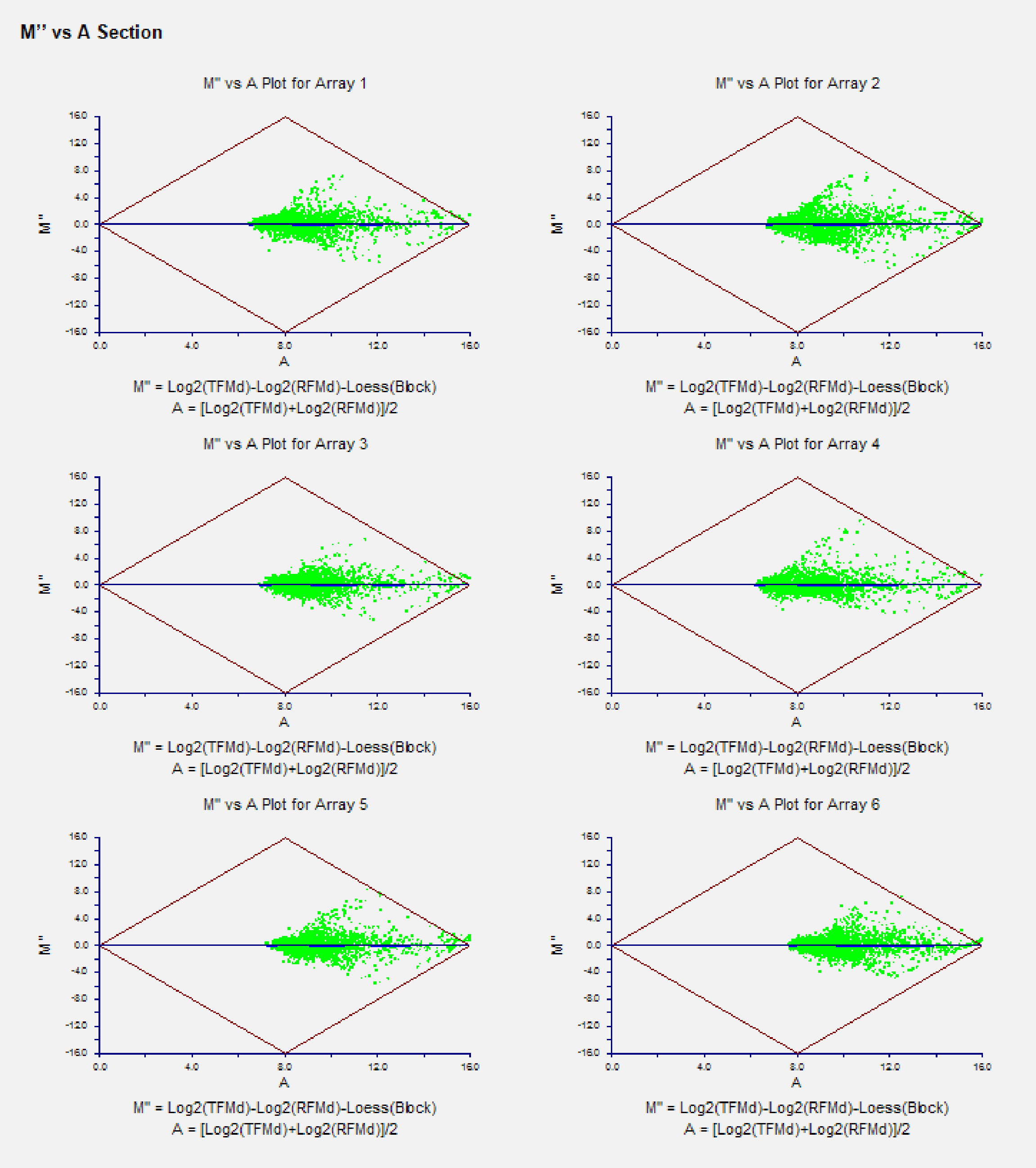
M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 6

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

These M vs A plots can be used to monitor dye bias. The M value (Y-axis) is the Target minus Reference difference in Log2(Foreground Median) intensities. The A value (X-axis) is the average of the Log2(Foreground Median) intensities. The diamond shows the physical limits of the M vs A coordinates when a 16-bit scanner is used. The loess line is overlaid. If there is no dye bias, the loess line will be near zero.

M’’ vs A Section



M’’ vs A Section

M" vs A Plot for Array 1

M" = Log2(TFMd)-Log2(RFMd)-Loess(Block)
A = [Log2(TFMd)+Log2(RFMd)]/2

M" vs A Plot for Array 2

M" = Log2(TFMd)-Log2(RFMd)-Loess(Block)
A = [Log2(TFMd)+Log2(RFMd)]/2

M" vs A Plot for Array 3

M" = Log2(TFMd)-Log2(RFMd)-Loess(Block)
A = [Log2(TFMd)+Log2(RFMd)]/2

M" vs A Plot for Array 4

M" = Log2(TFMd)-Log2(RFMd)-Loess(Block)
A = [Log2(TFMd)+Log2(RFMd)]/2

M" vs A Plot for Array 5

M" = Log2(TFMd)-Log2(RFMd)-Loess(Block)
A = [Log2(TFMd)+Log2(RFMd)]/2

M" vs A Plot for Array 6

M" = Log2(TFMd)-Log2(RFMd)-Loess(Block)
A = [Log2(TFMd)+Log2(RFMd)]/2

These M’’ vs A plots show the dye bias correction obtained when subtracting the block loess value. The loess line is overlaid, but is difficult to see since it is at or near the zero line, indicating the dye bias has been removed.

121-50 GenePix® GPR File Pre-Processing Engine

Chapter 130

Generic Two-Channel File Pre-Processing Engine

Obtaining Expression Values from Two-Channel Array Files

The main purpose of this chapter is to describe the process of obtaining relative expression values from two-channel files using the ***GESS*** Generic Two-Channel File Pre-Processing Engine. The input for this procedure is the summary files that are generated by image analysis software or similar means. The Generic Two-Channel File Pre-Processing Engine produces relative expression values for each gene that are then ready for statistical analysis. This procedure also gives a variety of filtering options as well as quality control and visualization tools.

The input files are assumed to contain a column of gene names, and a column for each of the channels, denoted Cy5 (Cyanine 5, red) and Cy3 (Cyanine 3, green). One channel files may also be read with this procedure, but another engine is dedicated to one-channel files. This engine accepts tab-, comma-, space-, and semicolon-delimited files. Column names or numbers may be used to identify the columns of interest. Columns of interest may be selected from a large number of columns in the file.

Following a brief background to the concept of microarrays, this chapter discusses many of the principles and practical aspects of two-channel arrays, including array production, image analysis, array and spot quality, filtering issues, and normalization. Reference and paired experimental designs are also presented as well as the dye-swap technique. The chapter concludes with a tutorial of the entire process of using the Generic Two-Channel File Pre-Processing Engine to obtain relative expression values from image-processed files.

Chapter Structure

Background

An overview of microarray concepts is presented first. This section is designed to familiarize a non-biologist with the concepts of DNA expression and microarray hybridization.

Nine Steps to Obtain Relative Expression Values

The background is followed by a summary of the nine steps required to obtain final relative expression values for a single ***two-channel*** array, which can then be used in comparison analysis.

Step 1 – Microarray Fabrication. Some common microarray fabrication methods are described.

Step 2 – Hybridization. Two samples are each labeled with a different dye, mixed, and then introduced onto the array. The sequences that are complementary to each probe will bind to the probe sequences. The two samples compete for binding at each probe.

Step 3 – Array Image Construction. The array is scanned twice, once for each dye, producing two image files. The image files contain a value (representing a shade of gray) for every pixel on the array.

Step 4 – Array Image Processing. Image processing software is used to locate each spot and separate foreground and background regions for both dyes.

Step 5 – Image Quantification. Image processing software produces summaries of the foreground and background pixels for both dyes at each spot. Relative intensities, comparing Cyanine 5 (Cy5, red) to green (Cyanine 3) dyes, are also produced.

Step 6 – Array Spot Types. Results from specially designed probes can be used to assess array quality.

Step 7 – Whole Array Quality. Specialized plots and numeric summaries of the whole array give indication of possible dye bias, spatial variation, or other artifacts that indicate whether values obtained from the array can be trusted for comparison.

Step 8 – Array Individual Spot Quality and Filtering. Pixel summaries and other values for each individual spot give indication of the spot quality. A variety of filters can be used to remove spots of questionable quality.

Step 9 – Array Normalization. A whole array normalization is recommended to correct for dye bias.

The result of these nine steps is a column of relative expression values that can be compared to corresponding values of other arrays in the experiment.

Two-Channel Image Processed Files

Common two-channel image processed files are described.

Two-Channel Designs

Paired and Reference Experimental Designs are described, as well as the dye-swap technique.

Entering Two-Channel Files

Details of entering two-channel files into the spreadsheet are explained.

Procedure Options

The options available in ***GESS*** for preprocessing two-channel files are described in detail.

Tutorial/Examples

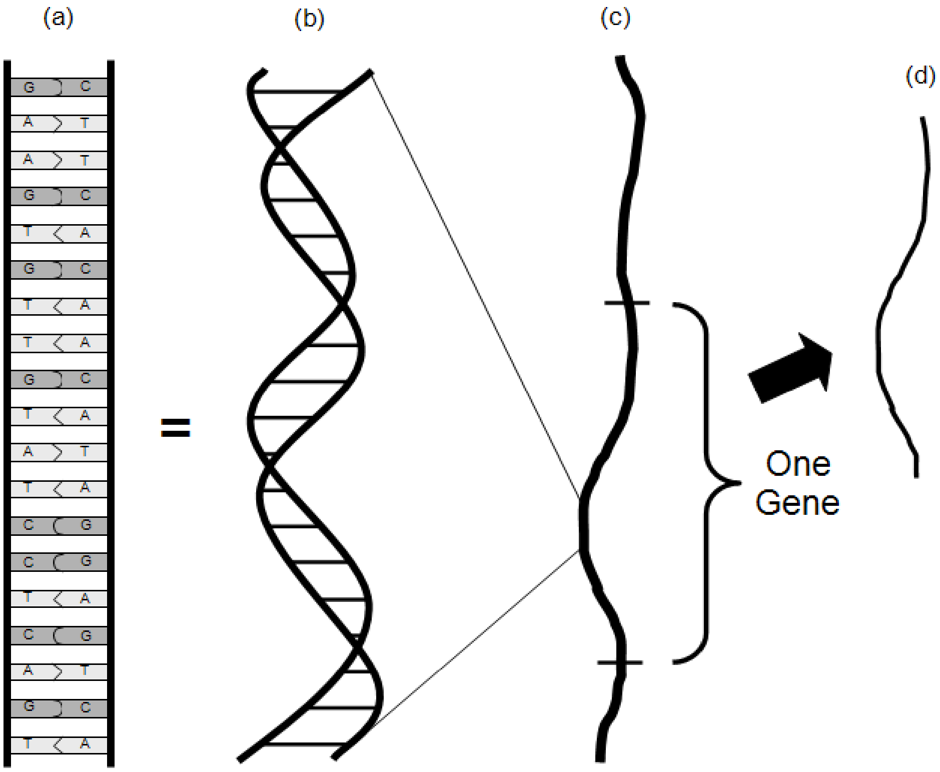
Examples of pre-processing two-channel files in ***GESS*** are shown.

Background

Gene Expression

The general process of gene expression in a cell begins with the DNA. Each DNA molecule has the well-known double helix design. Each rung or step of a DNA molecule is made up of two *nucleotides*. The two nucleotides bonded together are called *base pairs*. In DNA the 4 possible nucleotides are A, T, C, and G (for Adenine, Thymine, Cytosine, and Guanine, respectively). The nucleotide A can only form a base pair with T, and vice versa. Similarly, C can only bind to G, and vice versa. A gene is a unique segment of a DNA molecule consisting of a series of base pairs ranging from about 50 to thousands of base pairs in length. When the need for a specific protein in the cell is identified, the gene for that protein is “read” and a *messenger RNA* (mRNA) is produced in a process called *transcription*. mRNA molecules are single-stranded molecules which are essentially copies of the gene segment of one of the two DNA strands. The mRNA is then used to produce a protein that is specific to that mRNA molecule in a process called translation.

DNA overview. (a) A ladder representation of a DNA segment showing 20 complementary base pairs. (b) A drawing of the three-dimensional form of the corresponding double helix. (c) The 20 base pairs of (a) and (b) are only a small section of the total DNA double strand. A gene is a segment of the DNA helix that contains the code for the production of a protein. (d) A single stranded mRNA molecule is generated when the gene is expressed. The mRNA molecule will be used to produce a protein that is specific to the gene of (c).

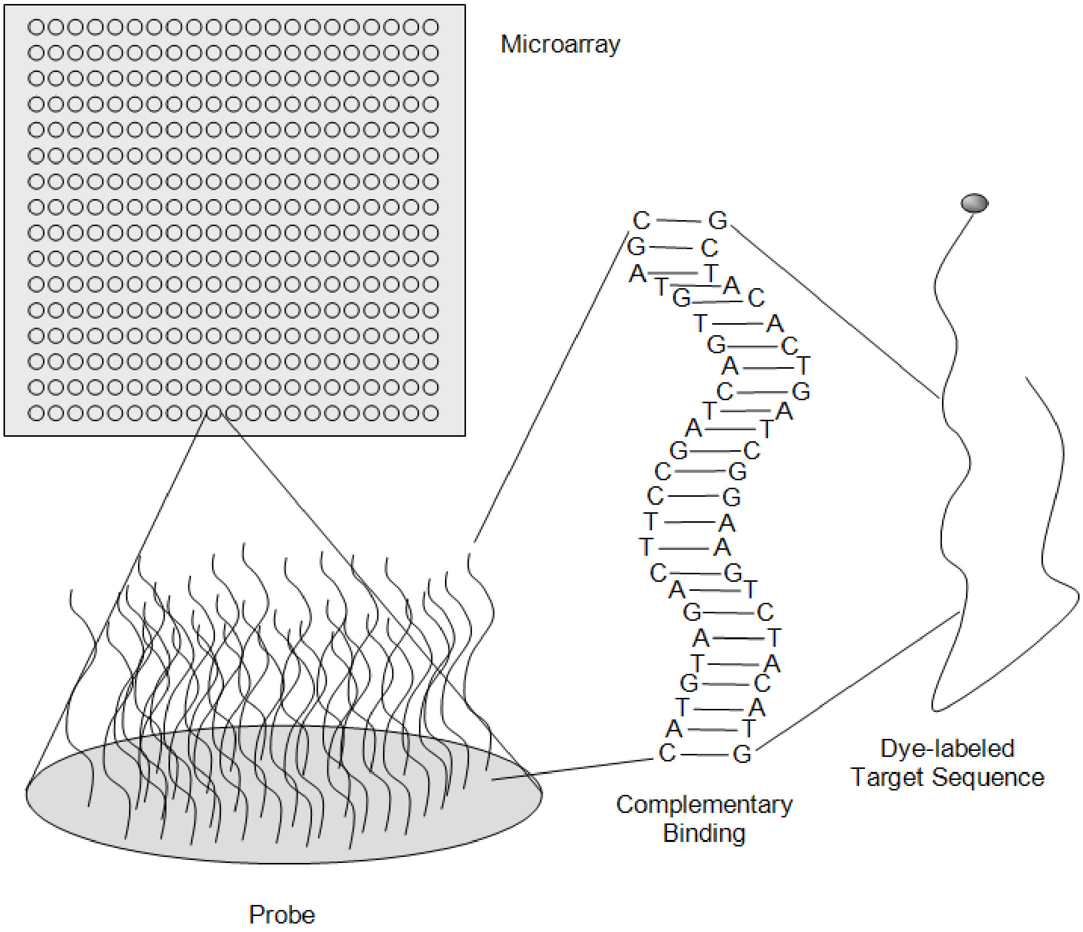


The newly created protein can then be used in the cell to perform the needed function. A gene that is in the process of producing or has produced a protein is said to be *expressed*. Expression of a given gene at a given time can thus be measured either by the amount of mRNA or protein (corresponding to that gene) in the cell. Microarrays are currently the prominent tool for quantifying the amount of mRNA in the cell (or collection of cells) for hundreds or thousands of genes simultaneously.

The Microarray

On a typical microarray, there are several thousand spots with *probes* (see below) of known identity, with each probe corresponding to a gene of interest. The probe sequences on each spot are designed to attract only sequences that are expressed by the gene to which that spot corresponds. A spot with the attached probe sequences may collectively be called a probe.

Below is a microarray drawing depicting the arrangement of spots on a spotted array (top left), the identical probe sequences on an individual spot (bottom left), and a segment of the probe for this spot that uniquely attracts the mRNA (or cDNA, a more stable mRNA replicate) strands of interest (center and right).



Hybridization (Binding to the Microarray)

The mRNA expressed in an experimental unit is obtained from some of the experimental unit's cells (i.e., blood or tissue), converted to cDNA (a nearly equivalent, but more stable molecule) using a process called reverse transcription, and labeled with fluorescent dye. When the solution containing the cDNA is exposed to a microarray, each of the cDNA sequences will bind to the probe sequences to which it complements. Thus, only sequences with perfect complementation along the entire sequence should bind to the corresponding probe (see figure above). The cDNA from genes which are expressed in higher quantities will hybridize (bind) to the corresponding probe in higher quantities. The amount of hybridized material for each spot can then be measured using the intensity of fluorescence from the bound cDNA when exposed to laser scanning. A scanner (or scanning machine) measures the intensity of fluorescence for every spot on the array. The result of a single microarray scan is several thousand intensities representing the amount of mRNA expression of those genes that are probed on the array.

Nine Steps to Obtain Relative Expression Values

Step 1 – Microarray Fabrication

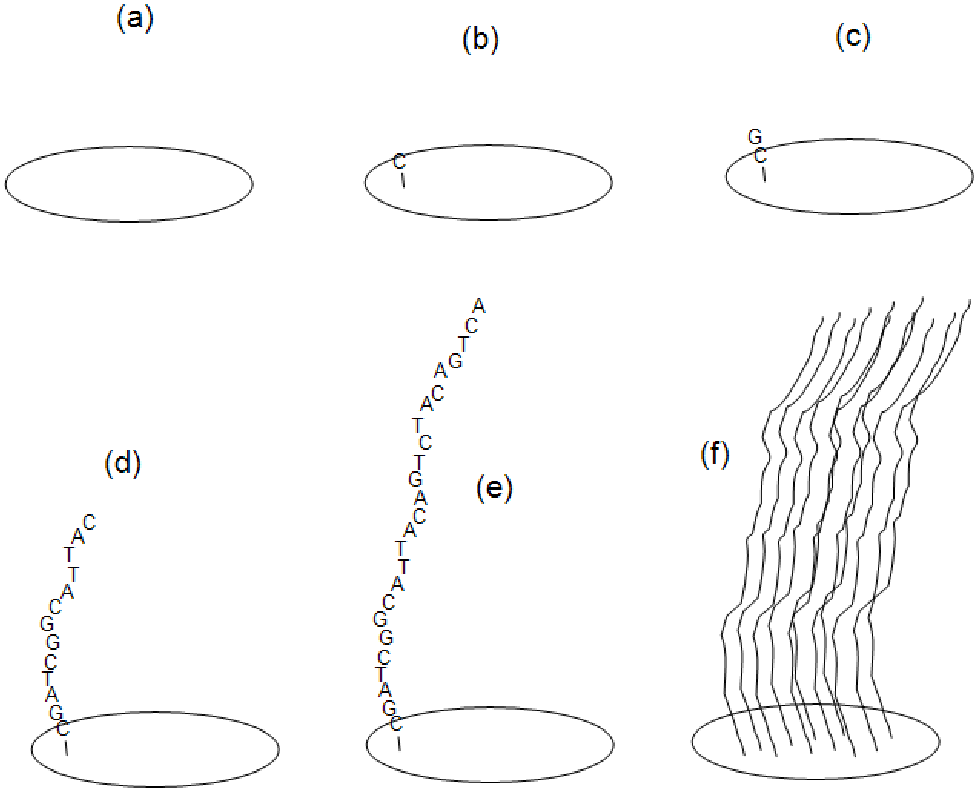
Two types of microarray fabrication that are commonly used in practice, *in situ* and spotting, are described below.

In Situ Microarray Fabrication

With *in situ* microarray synthesis, probe sequences are constructed nucleotide by nucleotide, directly on the array. A general advantage of this method of probe synthesis is that the sequence of every probe is known exactly. A disadvantage is that *in situ* synthesis techniques usually limit the probe sequences to lengths much shorter than those of spotted probes.

Two approaches are commonly used in *in situ* fabrication of two channel arrays: ink jet and electrochemical. In the ink jet printing process, four cartridges, each containing one of the four nucleotides, are used to deposit the nucleotides in the correct location, based on digital sequence files. In the electrochemical approach, small electrodes are used to guide the synthesis of the probe sequences.

The progress of generating probe sequences is seen in the figure. Multiple copies of the same probe sequence are generated in a tiny circular region to form a single probe spot. Thousands of probes (spots) are synthesized onto a single array.

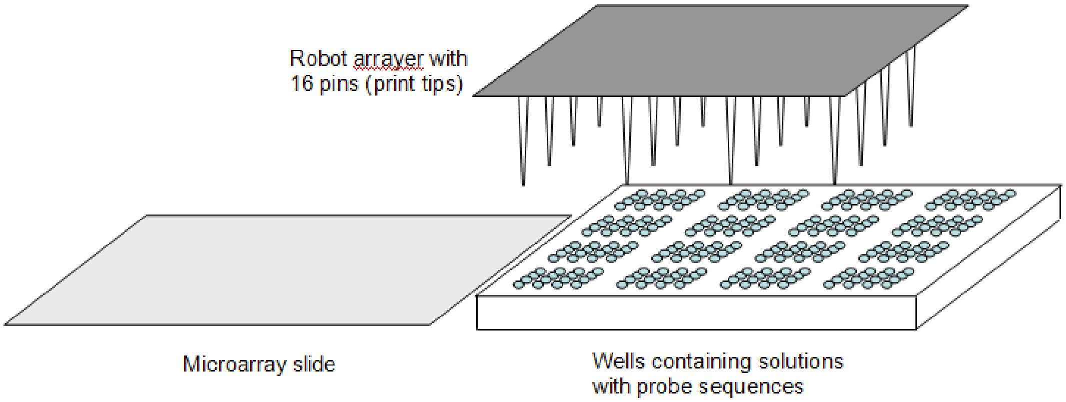

(a)
(b)
C
(c)
G
C
(d)
C A T T A C G G C T A G C
(e)
A C T G A C A T C T G A C A T T A C G G C T A G C
(f)

Spotted Microarray Fabrication

For spotted (or deposited) arrays, the probe sequences are prepared away from the chip and then spotted onto the slide in small quantities using robots with thin pins. The probe sequence solution for each probe is made by making many copies of a single known sequence using a technique called PCR. The robots dip the pins into each sequence solution and then touch the pins to the surface of the slide to form a spot of probe material (see below). Groups of spots that are printed with the same pin are called pin groups, or print-tip groups. Because print-tip groups can differ according to the characteristics of each pin, adjustments for differing print-tip groups are often made to expression values in a process called print-tip group normalization.

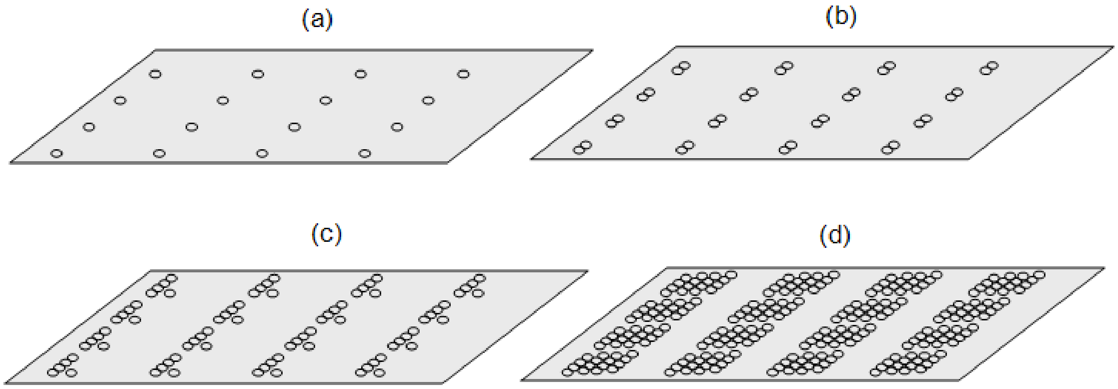
Robot Arrayer

Below is a drawing of a robot arrayer lifting probe solutions to be deposited on the microarray slide. Each well contains a solution with multiple copies of the same sequence. The sequence in each well is different from that of all other wells. The arrayer shifts down one well after each spotting.



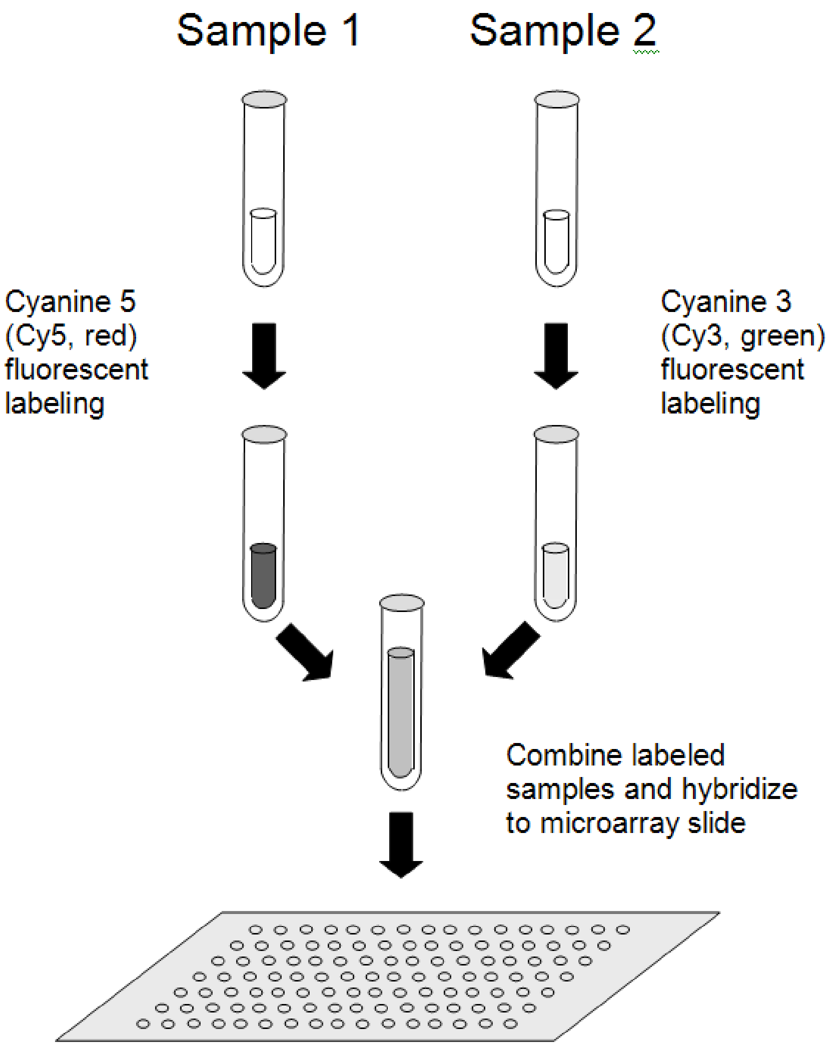
Progress of Fabrication of a Spotted Microarray

The figure below shows (a) the slide after the first group of spots is deposited, (b) the slide after the second spotting, (c) the slide after the fifth spotting, and (d) the slide after the final (16th) spotting. Each block of 16 spots is formed by the same pin and is called a pin group or print-tip group. A total of 16*16 = 256 different probes (representing, perhaps, 256 genes) have been deposited on the final microarray slide.



Step 2 – Hybridization

Two-channel microarrays refer to those for which two samples (and two dyes) are analyzed on each array. Two-channel microarrays are also called two-color microarrays. One cDNA sample is labeled with Cyanine 5 (Cy5, red) dye, and the other sample is labeled with Cyanine 3 (Cy3, green) dye. The samples are mixed and then introduced onto the array to compete for hybridization at each spot, as shown in the following diagram of the two-channel fluorescent labeling process.

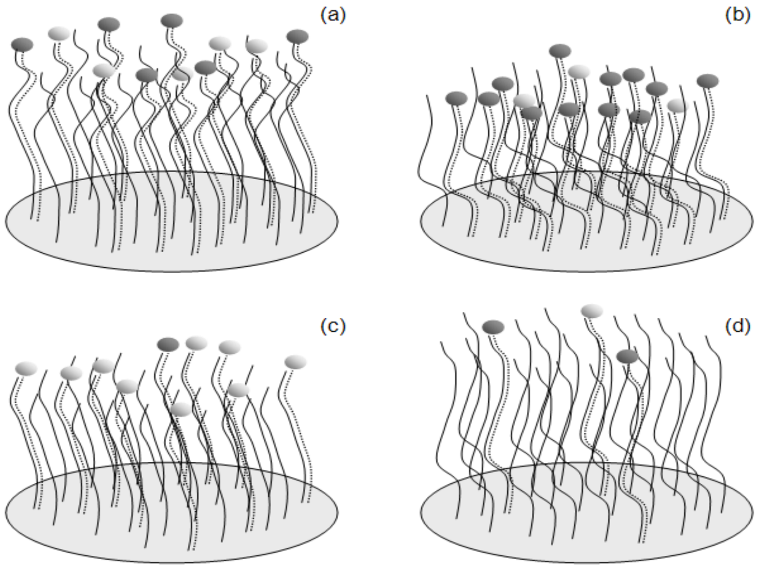

Sample 1
Sample 2
Cyanine 5 (Cy5, red) fluorescent labeling
Cyanine 3 (Cy3, green) fluorescent labeling
Combine labeled samples and hybridize to microarray slide

If a specific sequence is highly expressed in one of the samples (say, Sample 1) and has low expression in the other sample (say, Sample 2), the probe for that sequence should bind more Sample 1 sequences than Sample 2 sequences. This process is called competitive hybridization.

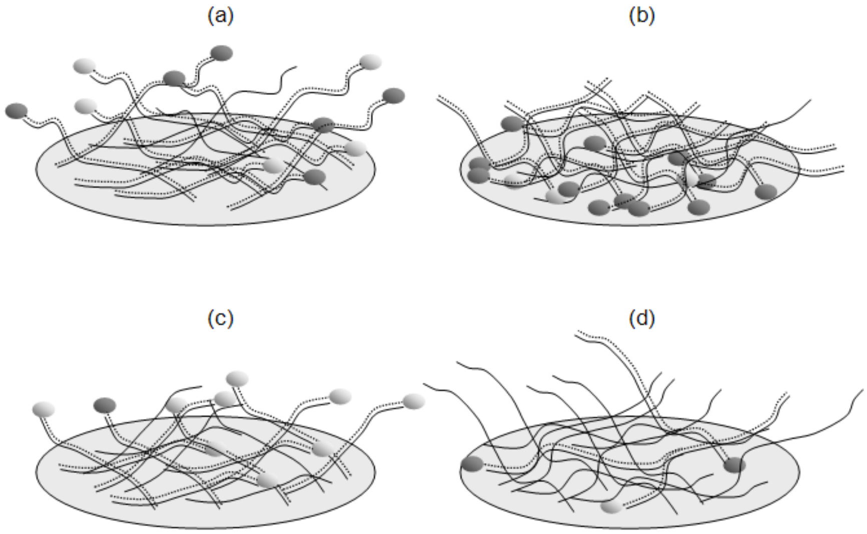
Competitive Hybridization

Below are some examples of competitive hybridization at individual spots. The dotted line sequences with dark dots attached represent Sample 1 cDNA with Cyanine 5 (Cy5, red) fluorescent labels. The dotted line sequences with light dots attached represent Sample 2 cDNA with Cyanine 3 (Cy3, green) fluorescent labels. Each of the four examples show varying amounts of competitive hybridization. For the probe in (a), there is nearly equal expression among both channels. In (b), there is high expression of this gene in Sample 1, low expression in Sample 2. In (c) can be seen very low expression of this gene in Sample 1, but high expression in Sample 2. There is very low expression for this gene in both samples in (d).

In Situ Synthesized Array Competitive Hybridization



Spotted Array Competitive Hybridization

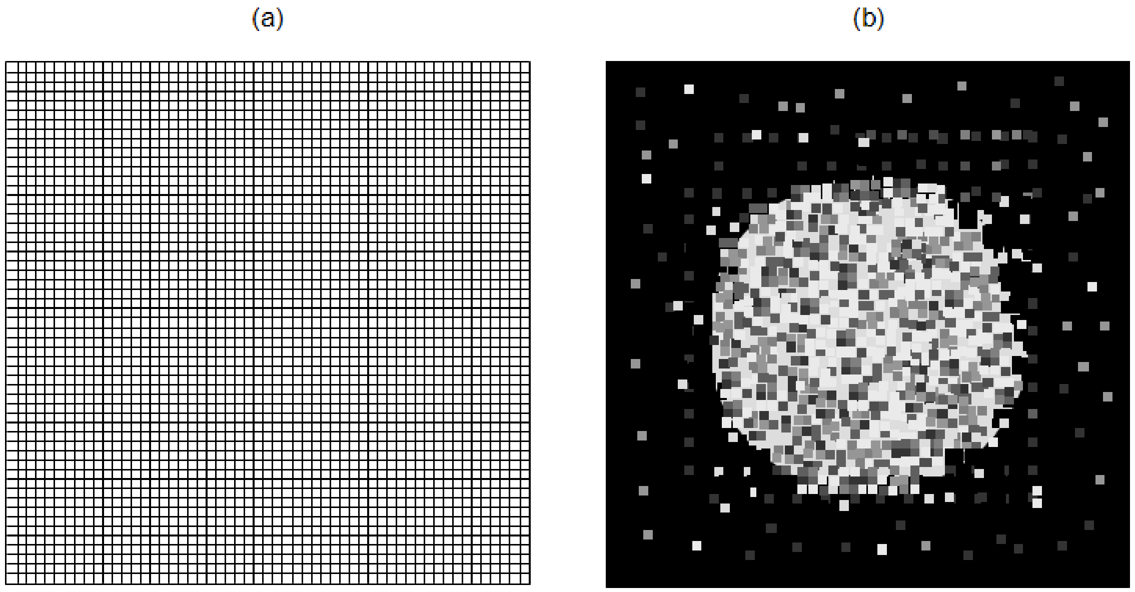


Step 3 – Array Image Construction

Following competitive hybridization, the laser of a scanning machine is used to illuminate the fluorescent dye of one of the channels (e.g., Cyanine 5) across the whole array, creating a high resolution black-and-white image for that channel. The frequency of the laser is then adjusted (or a different laser is used) to illuminate the fluorescent dye of the other channel (e.g., Cyanine 3), creating a second black-and-white image. Each of the images is usually stored as Tag Image File Format (.tif) file. The image is made up of a grid of pixels. Because each pixel is stored using 16 bits of memory, each pixel can take on any of 2^{16} = 65,536 shades of gray. The numeric range for each pixel is thus 0 to 65,535. The number of pixels in each spot depends upon the resolution (total number of pixels) of the image and the size of the spot (see the figure below).

Pixel Grid and .tif Image following Laser Scanning

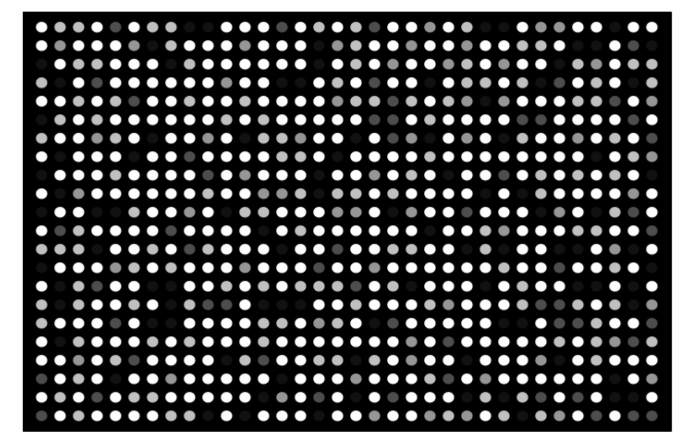
A pixel grid for the region of a single probe spot is shown in (a). A drawing of a .tif image following laser scanning is shown in (b). The number of pixels in a spot may range from below 50 to several hundred, depending upon the size of the spot and the resolution of the image.



Artificial colors (commonly, red and green) are often used in image construction software to visualize the expression represented on each array. The colors may be superimposed to form a single array image that displays the relative illumination at each spot. On the superimposed image, red spots are often used to indicate the Cyanine 5 channel fluoresced more than the Cyanine 3 channel. Green spots indicate the Cyanine 3 channel fluoresced more than the Cyanine 5 channel. Yellow spots indicate equivalent (or nearly equivalent) fluorescence. Brighter spots reveal probes that attracted larger amounts of labeled cDNA. Dimmer (darker) spots indicate smaller amounts of labeled cDNA were bound (see the figure on the following page).

Intensity Image Drawing

Below is a drawing of an artificial superimposed intensity image similar to that output in image construction software. Bright spots indicate a large amount of cDNA hybridization. Dark spots reflect low or no hybridization. If red, yellow, and green colors could be seen, they would indicate *relative* hybridization at each spot.

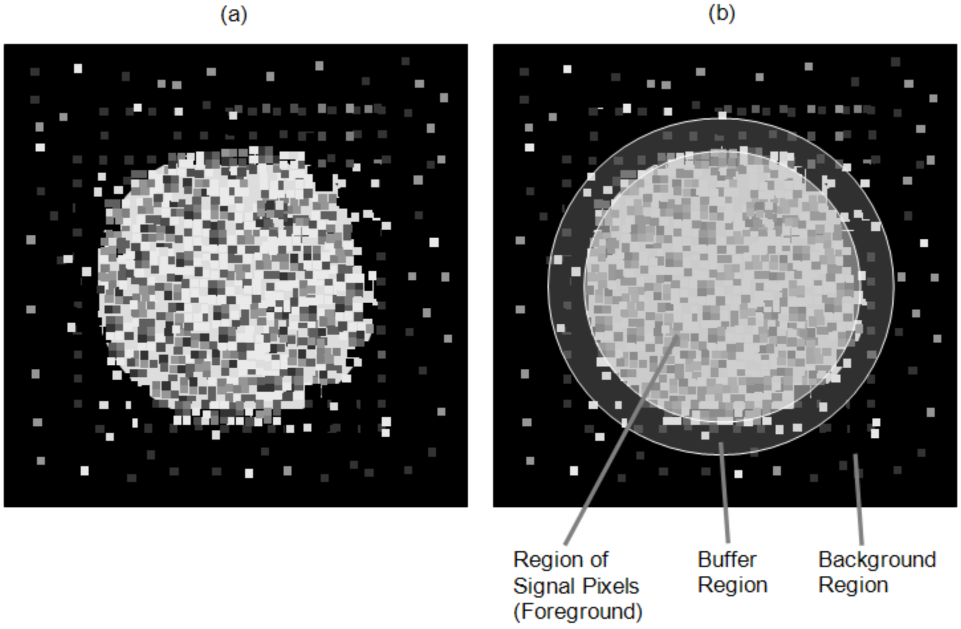


Step 4 – Array Image Processing

There are two main steps in image processing. First, the location of each spot is determined (spot finding or gridding). Second, the pixels of each spot are separated into foreground/signal and background regions (image segmentation). It is intended that the foreground region corresponds to the region where the probe is located, while the background region should be a region where no probe sequences were positioned. Special algorithms, which are continuously being improved, are used in image processing software to perform both of these steps. A properly found and segmented image should look something like (b) in the figure below.

Image Segmentation

The intensity image before spot location and segmentation is seen in (a). The intensity image after spot location and segmentation is shown in (b). A buffer region is used to assure pixels on the border are not misclassified.

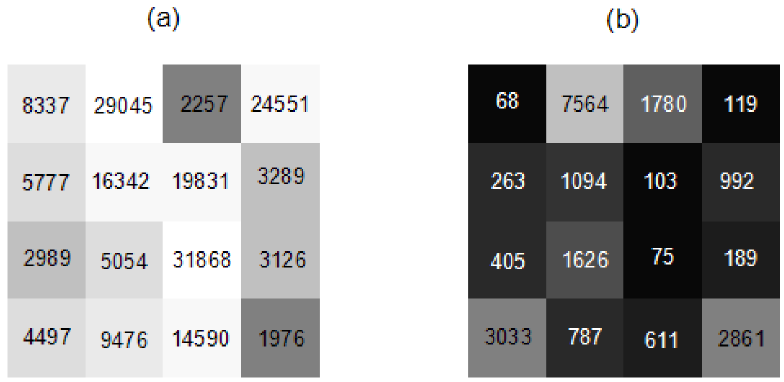


Step 5 – Image Quantification

Summaries of foreground and background pixel regions for each channel that are commonly provided by image analysis software are the mean, median, standard deviation, and total intensity, among others. The mean or median are usually used as final intensity measures for each spot of each channel. The other measures are used for inspecting spot or array quality or for filtering undesired spots.

Pixel Intensities

The figure that follows shows (a) sixteen pixels from the foreground region of a scanned spot with associated numeric intensities, and (b) sixteen pixels from the background region of a scanned spot with associated numeric intensities. Summaries of these numeric intensities (e.g., mean, median, and standard deviation) are obtained with image analysis software.



(a)

8337	29045	2257	24551
5777	16342	19831	3289
2989	5054	31868	3126
4497	9476	14590	1976

(b)

68	7564	1780	119
263	1094	103	992
405	1626	75	189
3033	787	611	2861

In a reference design, which is described in detail later in the chapter, the two samples to be analyzed in a single array are designated *target* and *reference* samples. The target sample is labeled with either the Cyanine 5 (red) or the Cyanine 3 (green) dye. The reference sample is labeled with the other dye. The goal of image quantification is obtaining a single value that reflects the relative expression of the target to reference channels at each spot. Two common measures of the relative expression of the two channels are the intensity difference (*Target – Reference*), and the log of the intensity ratio, *M*,

$$M = \log_2(Target / Reference) = \log_2(Target) - \log_2(Reference),$$

where *Target* and *Reference* represent, for example, the median foreground intensity for the target and reference channels, respectively. A common measure of overall brightness for each spot is

$$A = \log_2 \sqrt{Target * Reference} = \frac{\log_2(Target) + \log_2(Reference)}{2}$$

M and A are mnemonics for *m*inus and *a*dd (or *a*verage, or *a*bundance), respectively.

Using Logarithms

Dudoit et. al (2002) suggest using logged intensities rather than absolute intensities for the following reasons: “(i) the variation of logged intensities and ratios of intensities is less dependent on absolute magnitude; (ii) normalization is usually additive for logged intensities; (iii) taking logs evens out highly skewed distributions; and (iv) taking logs gives a more realistic sense of variation.” They also note that “logarithms base 2 are used instead of natural or decimal logarithms as intensities are typically integers between 0 and 2^{16} – 1.”

Step 6 – Array Spot Types

Although the majority of spots on an array will be used to compare gene expression levels across treatments, some arrays contain spots that are used only for array quality purposes. Below is a description of commonly used types of array quality spots.

Positive (Calibration) Control Spots: Spots containing probes corresponding to genes that are known to be expressed in all cells of the type under consideration. In a two-channel system, it is expected that both channels of positive control spots will have high (well above background) and equal expression.

Negative control spots: Spots containing probes that are known to be *not* expressed in cells of the type under consideration. In a two-channel system, it is expected that both channels of negative control spots will have low (near background) and equal expression.

Spike-in (Ratio) Control Spots: Spots containing probes that hybridize to cDNA that is entered into the cDNA solution in known quantities and proportions. These differ from positive controls in that the cDNA is introduced directly into the hybridizing solution, not expressed in the cells. These spots usually correspond to genes that are known to be *not* expressed in cells of the type under consideration. In a two-channel system, a variety of spike-in control spots are often used to exhibit varying amounts of *relative* expression at varying intensities. For example, one spot may correspond to a 3-to-1 red-to-green ratio, while another spot may designate a 1-to-3 red to green ratio.

Blank (Empty) Spots: Spots that contain no probe sequence and are spotted with water only or with nothing. They are used to estimate background hybridization levels.

Buffer (Empty) Spots: Spots that contain no probe sequence and are spotted only with buffer solution. They are used to estimate background hybridization levels.

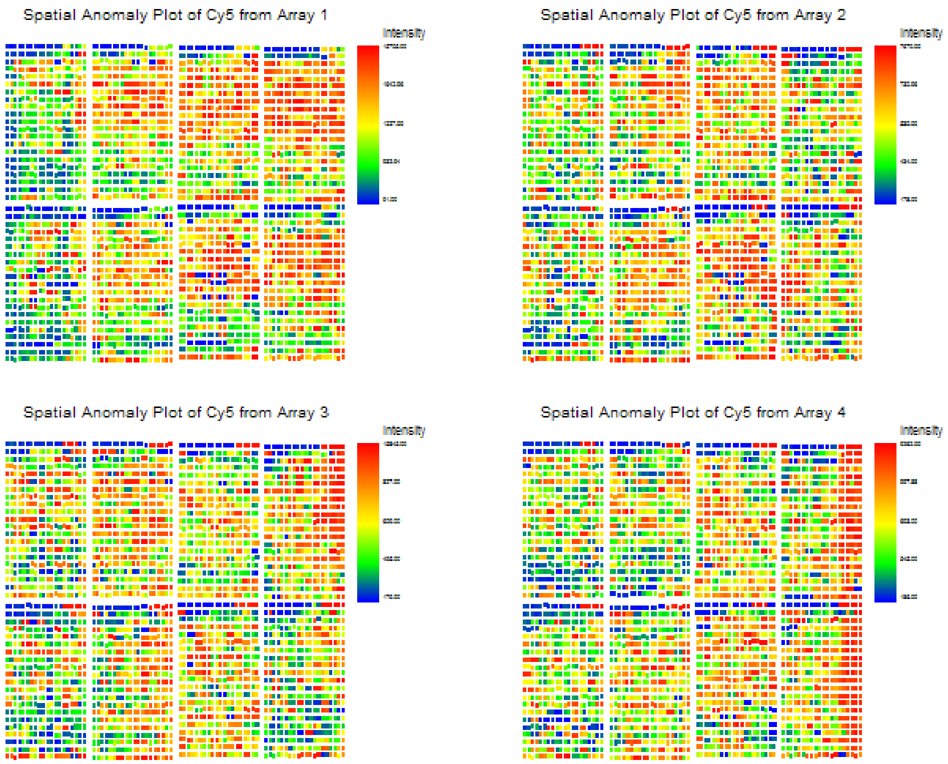
Step 7 – Whole Array Quality

Each microarray should be examined to assure the expression values of the array as a whole can be compared to the corresponding values of other arrays. The following can be used to determine microarrays of questionable quality.

Spatial Anomaly Plot

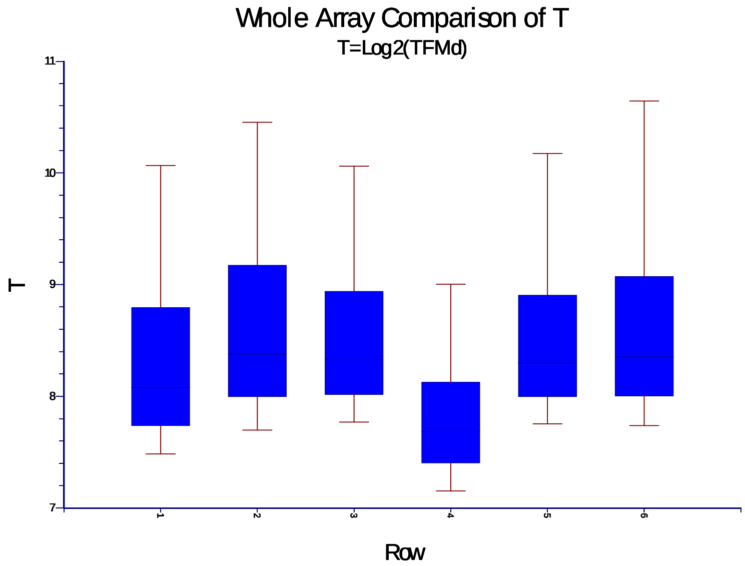
A spatial anomaly plot is a reconstructed picture of the whole array where intensities are represented by a color spectrum. If there are no anomalies, the distribution of color should be evenly dispersed throughout the plot. The following are four spatial anomaly plots of the median foreground intensity of the Cyanine 5 (Red) channel.

Spatial Anomaly Plots - Cy5



Array Comparison Box Plot

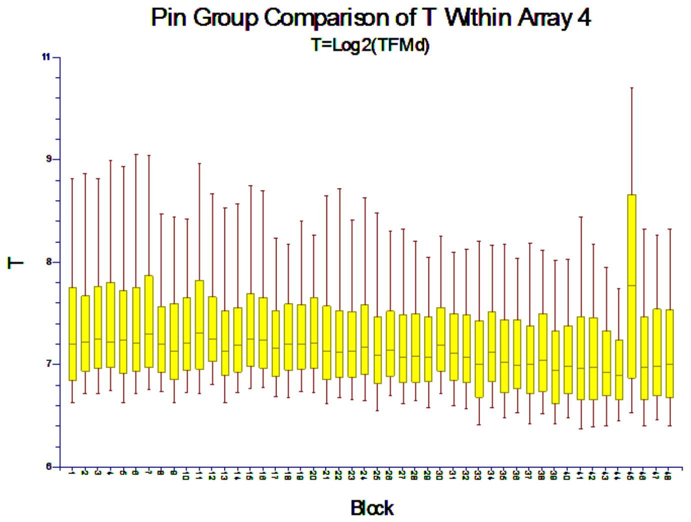
All arrays in the experiment may be compared for a given summary measure using a single side-by-side box plot graph. This example compares the Log2(median foreground intensities) of the Target sample of six arrays.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Pin (Print-tip) Group Box Plots

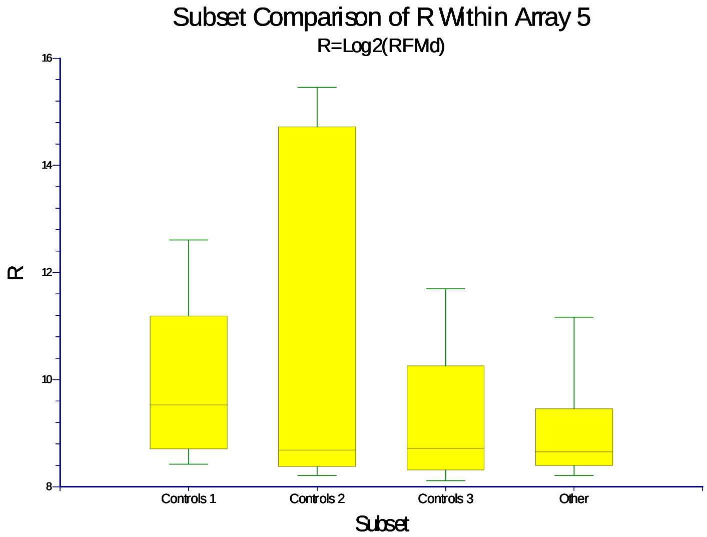
If blocks or pin (print-tip) groups exist on the array, spatial bias may also be seen by looking at box plots of the intensities of each of the blocks or pin (print-tip) groups. If one box plot is quite different from the others, it signals the print-tip (pin) used may be different from the others. The figure shows the box plots of Log2(median foreground intensities) of the Target sample. Pin (print-tip) group 45 may be suspect on this array.



Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Subset Box Plot

The intensity values of subsets can be compared to other subsets and non-subset spots using a subset side-by-side box plot. Often the subsets are various controls. The subset box plot below shows summaries of the Log2(reference foreground medians) for 3 controls and all other genes (spots).

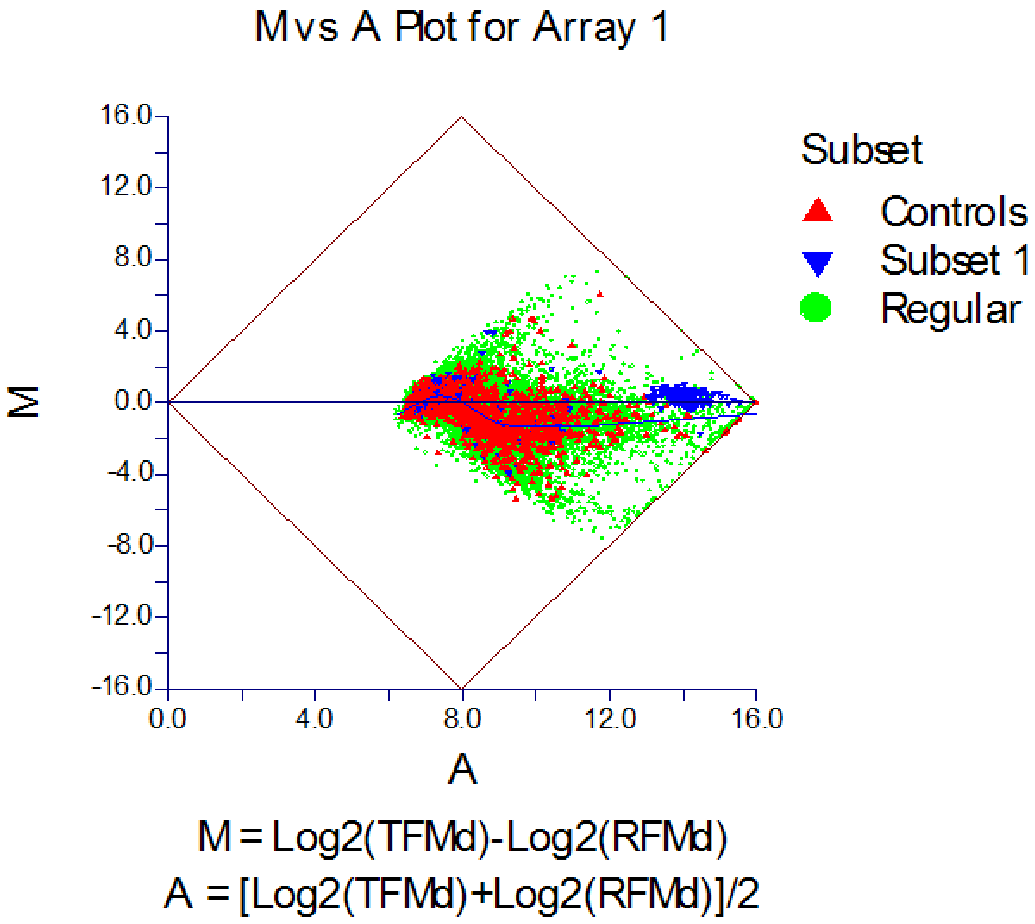


Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

M vs A Plot

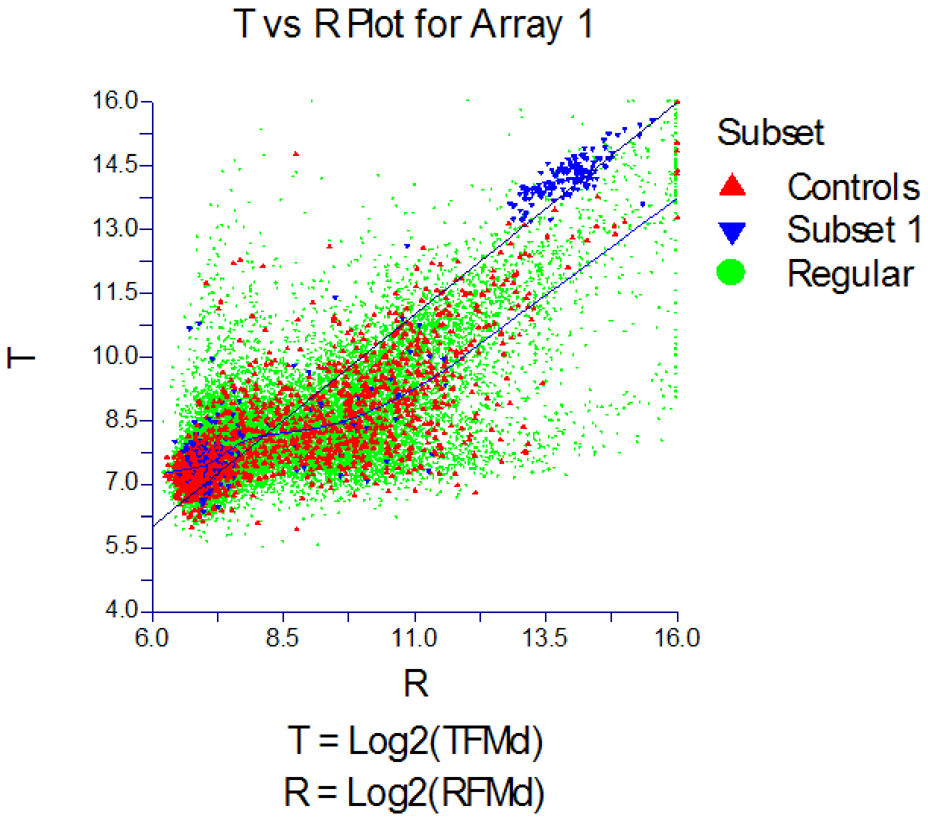
The M vs A plot is a scatterplot with a relative intensity measure (M) on the Y axis, and average intensity (A) on the X axis. This plot is useful for visualizing the relationship between dye-bias and intensity. Dye-bias occurs when one of the dyes produces a brighter image than the other dye for reasons other than differential expression. If there is no dye-bias, M values should be centered near zero across all values of A. If dye-bias exists, a banana shape is often the result. Plotting a loess line on the MA plot can be useful for viewing dye bias. A loess line is a specialized robust moving average that uses locally weighted polynomial regression. If the loess line is far from the zero line, there is evidence of dye bias. If the loess line is near the zero line, little or no dye-bias exists. An M vs A plot following loess normalization can be used to determine if the dye-bias problem has been properly corrected.

If control spots are shown in different colors, the M vs A plot is also useful for displaying how expression levels of control spots compare to those of the spots of interest. The diamond shows the physical limits of the M vs A coordinates when a 16-bit scanner is used. The overlaid loess line shows minor dye bias.



T vs R Plot

The relative intensity values of target and reference samples can be seen by plotting the target intensity measure on the Y axis, and the reference intensity measure on the X axis. This plot is similar to a rotation of the MA plot. This T vs R plot displays relative median foreground expression of target to reference samples. The upper physical limits of 16 can be seen here.



Numeric Summaries

Mean and standard deviation summaries for mean, median, or standard deviation of foreground and background regions of the spots are useful for comparing slides. Filter summaries across slides may also be good indicators of slides of questionable quality.

Step 8 – Array Individual Spot Quality and Filtering

Some useful indicators of individual spot quality include:

Standard deviation of Foreground or Background Pixels: This summary (of each channel) gives an idea of the uniformity of expression across the spot. Large standard deviations indicate non-uniformity. A large standard deviation can be determined by examining the whole slide numeric summary of standard deviations.

Flags: Some image processing software packages generate flags for spots which fail to meet some criteria. The criteria should be understood before using the flags as indicators of spot quality.

Saturation: Since each pixel has a limiting value of 65,535, pixels with such a value should be considered right-censored (unknown, but larger than 65,535). Spots with large proportions of saturated pixels may be termed saturated spots.

Weak Signal: Spots with low intensities are often near, or even below, the expression intensity of the background region. When expression intensities are near the background, it is difficult or impossible to estimate reliably the true expression level of that gene. It is only known that the gene is not highly or medium expressed. It is not known, however, whether the gene is expressed at very low levels, or not at all. Weak signal spots are essentially left-censored.

Weak Signal Considerations

Another aspect of foreground to background comparison should be mentioned here. Because the background is devoid of probe, it is not subject to nonspecific binding, as is the foreground, and thus may not accurately reflect the baseline it is intended to represent.

Many filters can be designed to remove values of questionable individual spot quality based on some cut-off value. However, it is much easier to flag a spot as one with questionable quality than it is to justify removal of the spot from future analysis. Spot filtering should not be treated lightly, as the following discussion illustrates.

An important decision that may have heavy bearing on the final analysis results is how the spots with low intensities will be treated. When the expression level of either channel is unknown, the relative expression of the two channels is also unknown. For example, suppose that for the Cyanine 3 (Cy3, green) channel at a given spot the estimated foreground is 100 and the estimated background is 150. The true expression intensity could be anywhere from 0 to 150 or more, but the techniques being used are not sensitive enough to distinguish a reliable estimate, due to background noise. Suppose further that the Cyanine 5 (Cy5, red) channel yields estimates of 600 for the foreground and 100 for the background intensity. Because 600 is well above 100, it can be assumed that the Cyanine 5 channel estimate is a reliable one. The difficulty in this situation lies in determining a good estimate of Cyanine 5 to Cyanine 3 relative intensity. The range of ratios is 600/150 = 4 to 600/0 = infinite. Even if logs are used, the ratio chosen will greatly affect the ensuing analysis, as will simply throwing the spot out, altogether.

Consider the following scenario in which the researcher throws out all weak-signal (or low relative intensity) spots. One thousand genes are to be compared based on 10 individuals each from a disease group and a non-disease group. It is of interest to determine which among the 1,000 genes is differentially expressed. Suppose there are four genes, which, unknown to the researcher, are highly expressed in the disease group, but not expressed in the non-disease group. The microarray experiment is run and the expression values for each of those four genes for all 10 individuals in the disease group are well above the background. However, the expression levels for those same four genes for the non-disease group are near background levels. Because the expression levels are near the background, the non-disease expression values for these four genes are all filtered from further analysis. A two-sample *t*-test is then attempted for each of the 1,000 genes. Unfortunately, the *t*-test cannot be run for the four genes of true differential expression, since for each of these four genes there are no expression values in the non-disease group. No evidence of differential expression is reported, and the four genes go undetected.

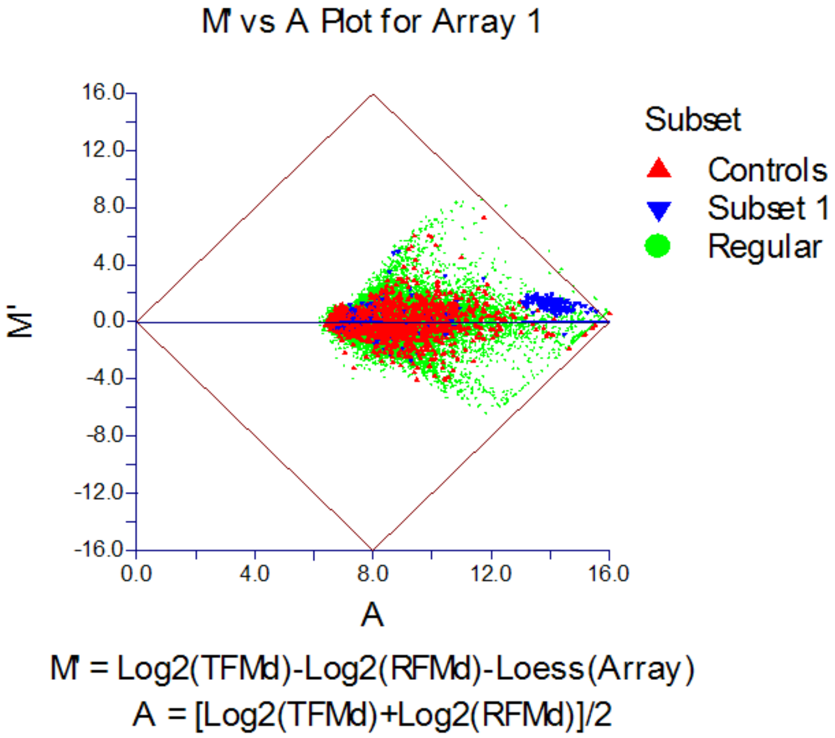
The preceding example reveals three aspects that should be considered when determining whether or not to filter a low intensity spot: (1) The foreground relative to the background for each dye individually; (2) The relative intensities of the two dyes at that spot; (3) The effect of removing that spot from the analysis, *across arrays*.

Step 9 – Array Normalization

The purpose of two-channel array normalization is correcting for dye-bias. Dye-bias occurs when one of the dyes produces a brighter image than the other dye for reasons other than differential expression. Dye-bias is usually dependent on spot intensity and may be viewed using an MA plot with the loess line overlaid (see *MA Plot* under the heading *Two-Channel Whole Array Quality* above). To correct for intensity dependent dye bias at each *A* value, the loess value is subtracted from each *M* value at that location. The *M* values become normalized *M* values. This normalization causes the *M* values to be centered at zero along the horizontal axis (see the figure below) and the dye-bias has been removed.

M vs A Plot Following Array Normalization

Below is the M vs A plot of the same data used for the M vs A plot shown previously, but here following whole array loess normalization. The overlaid loess line follows the zero line, indicating the dye bias has been removed. Some points may go beyond the physical limits of the diamond following loess normalization.



Two-Channel Image Processed Files

Image analysis software is often purchased with a scanner (or scanning machine) to summarize the pixel intensities of the scanned .tif image. Usually the image analysis software will implement a set of feature-finding algorithms, which are used to find and align each spot on the scanned array. Foreground and background regions are also determined. The foreground and background pixel intensities of each channel are summarized in a file that corresponds to a single array. Several files are generated for an experiment of multiple arrays.

The format of the image summary files varies according the scanner used and the image analysis software used. Two examples of such formats are described in the GenePix® and Agilent® Pre-Processing Engine chapters. The Two-Channel Pre-Processing Engine can be used to pre-process files of these and most other types. The general format need only be files that contain a column with a gene identifier and two columns that summarize the expression intensity of the two channels (e.g., Cy5 and Cy3, or Red and Green).

Data columns that are commonly found in image processed files and that may be used in the ***GESS*** Two-Channel Pre-Processing Engine are described below.

Gene Name The gene name or identifier for each gene.

Block (Pin Group) A number identifying the block or pin (print-tip) group of the probe/spot on the array.

X Coordinate The horizontal coordinate of the center of the spot.

Y Coordinate The vertical coordinate of the center of the spot.

Red (Cy5) F Median The median foreground pixel intensity for the Cyanine 5 (Cy5, Red) channel.

Red (Cy5) F Mean The mean foreground pixel intensity for the Cyanine 5 (Cy5, Red) channel.

Red (Cy5) F SD The standard deviation of the foreground pixel intensities for the Cyanine 5 (Cy5, Red) channel.

Red (Cy5) B Median The median background pixel intensity for the Cyanine 5 (Cy5, Red) channel.

Red (Cy5) B Mean The mean background pixel intensity for the Cyanine 5 (Cy5, Red) channel.

Red (Cy5) B SD The standard deviation of the background pixel intensities for the Cyanine 5 (Cy5, Red) channel.

Green (Cy3) F Median The median foreground pixel intensity for the Cyanine 3 (Cy3, Green) channel.

Green (Cy3) F Mean The mean foreground pixel intensity for the Cyanine 3 (Cy3, Green) channel.

Green (Cy3) F SD The standard deviation of the foreground pixel intensities for the Cyanine 3 (Cy3, Green) channel.

Green (Cy3) B Median The median background pixel intensity for the Cyanine 3 (Cy3, Green) channel.

Green (Cy3) B Mean The mean background pixel intensity for the Cyanine 3 (Cy3, Green) channel.

Green (Cy3) B SD The standard deviation of the background pixel intensities for the Cyanine 3 (Cy3, Green) channel.

Two-Channel Designs

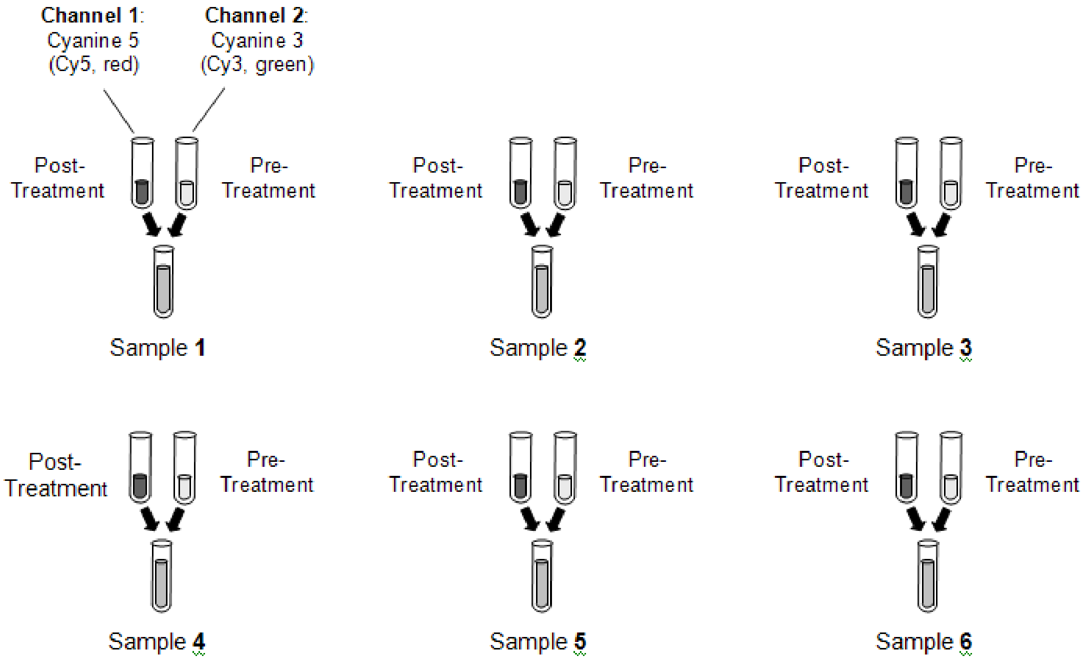
Two experimental designs may be used when using two channel microarrays: paired designs and reference designs.

Paired Design

The paired design is often used in two-channel experiments when the gene expression comparison to be made involves a natural pairing of experimental units.

As an example, suppose 6 cell samples are available for comparison. A portion of each of the 6 cell samples (before treatment) is reserved as a control. The same treatment is then given to each of the 6 remaining portions of the samples. It is of interest to determine the genes that are differentially expressed when the treatment is given. In this scenario, there is a natural before/after treatment pairing for each sample. The reserved control portions of each sample are labeled with Cyanine 3 (Cy3, green) dye, while the treatment portions are labeled with Cyanine 5 (Cy5, red) dye. From each sample, the labeled control and the labeled treatment portions are mixed and exposed to an array. The control and treatment portions compete to bind at each spot. The expression of treatment and control samples for each gene is measured with laser scanning. A pre-processing procedure is then used to obtain expression difference values for each gene. In this example, the result is 6 relative expression values (e.g., $\mathrm{Log}_2(\mathit{Post} / \mathit{Pre})$) for each gene represented on the arrays.

Paired Design, Six Arrays



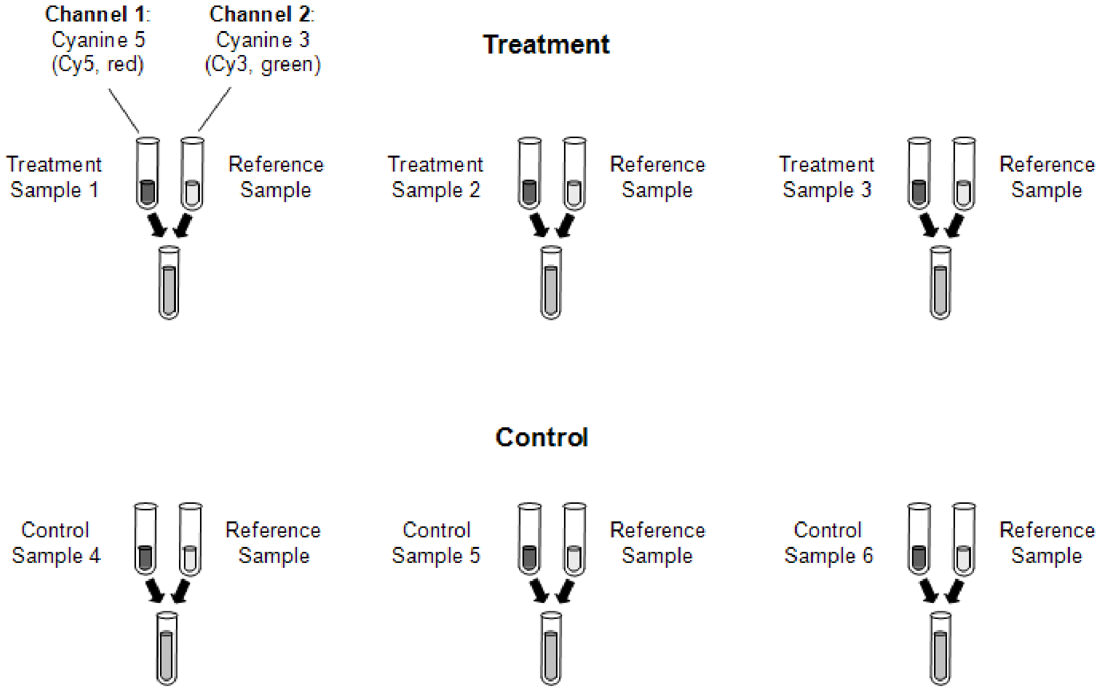
Reference Design

A two-sample reference design, or common reference design, employs an outside source of cDNA that is used as a reference for all samples in the experiment. Reference cDNA may be purchased separately or may be a combination of all cDNAs in the compared samples (The pros and cons of choice of reference cDNA is beyond the scope of this manual).

Suppose a treatment and control are to be compared. One group of experimental units serves as the control group. The other group of experimental units receives the treatment. Following treatment, cDNA is isolated for each of the experimental units. The cDNA for the treatment and control groups may be termed target cDNA. The target cDNA from both groups is labeled with Cyanine 5 (Cy5, red) dye. An outside source of cDNA, with (hopefully) most genes of interest expressed, is labeled with Cyanine 3 (Cy3, green) dye. This cDNA is the common reference, and is used as a baseline for all arrays of both groups. The intensity value for each gene of each array is the relative expression of the target cDNA to the reference cDNA at each spot (see data examples in the tables that follow).

The goal of the reference cDNA is to remove additional variation that may have been introduced in the experimental procedure. Array differences may be particularly pronounced when large periods of time pass between array hybridizations of a single experiment. Reference designs may also be employed in repeated measures/time-course designs.

Two-Sample Reference Design, Six Arrays

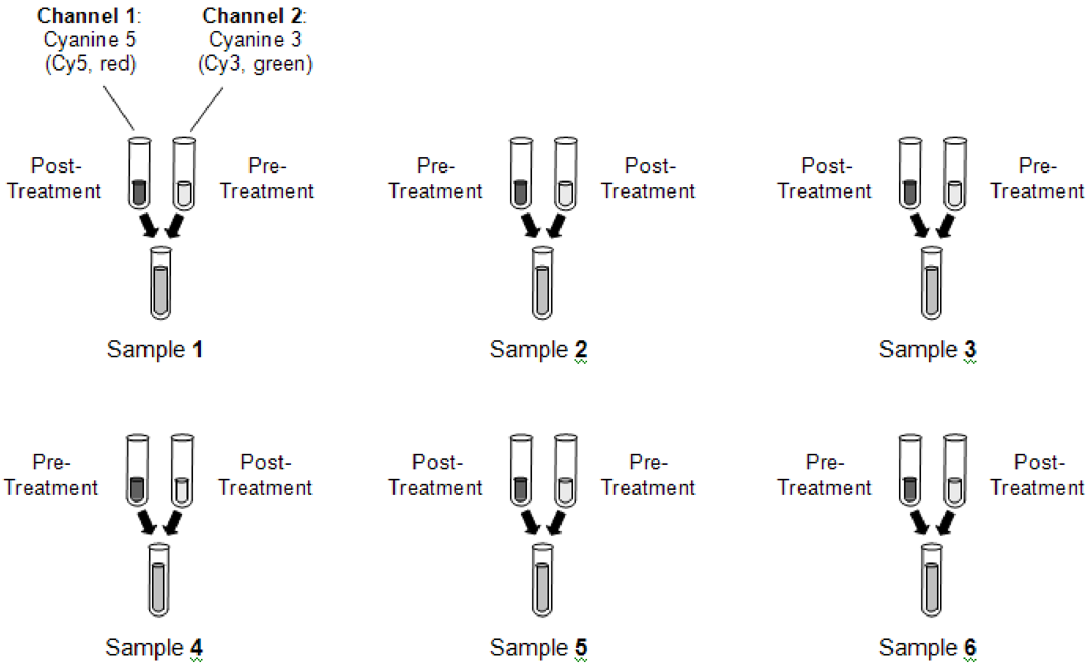


Dye Swap

Dye swap is a technique that may be employed in either paired or reference designs. The purpose is to remove systematic bias of the dye. To use this technique, the dye used is switched for a subset of the experimental units. For example, in the paired design example above, all 6 control portions are labeled with Cyanine 3 (Cy3, green) dye, while the 6 treatment portions are all labeled with Cyanine 5 (Cy5, red) dye. The dye swap technique could be employed by labeling half of the 6 controls with Cyanine 3 (Cy3, green) dye and the other 3 with Cyanine 5 (Cy5, red) dye. The 6 treatment portions would be labeled with the complement dye to that of the corresponding control portions. Careful record should be kept of which dyes are used on each array when performing an experiment with dye-swapping.

Dye-swap - Paired Design

Each of the 6 samples is divided into two portions. One portion serves as control. The other portion receives the treatment. Three of the treatment portions are labeled with Cyanine 5 dye. The corresponding control portions are labeled with Cyanine 3 dye. The other three treatment portions are labeled with Cyanine 3 dye. The corresponding control portions are labeled with Cyanine 5 dye. The samples are combined and then introduced onto a microarray slide. Six relative expression values are obtained for each probe: three are $\text{Log}_2(Cy5 / Cy3)$, the other three are $\text{Log}_2(Cy3 / Cy5)$.



Entering Two-Channel Files

This section describes how file names are entered into the spreadsheet in preparation for preprocessing. Two variables (columns) are required to run two-channel pre-processing, and a third is required to obtain output (.ges) files. Any name (entered by clicking on the Variable Info tab at the bottom of the spreadsheet) may be used for these variables.

Two-Channel File Name Variable

The two-channel file name variable is a column on the spreadsheet containing a list of paths and filenames of the files (previously generated by image analysis software) that are to be pre-processed. The files may be in different folders, but must all contain results for the same list of genes, and must have the same format (i.e., the expression values must start on the same row in each file and the columns must be arranged the same order in all files). This variable is required to run the Generic Two-Channel File Pre-Processing Engine procedure.

Target Sample Dye Variable

When two-sample (two-channel) microarrays are used, one sample may be termed the target sample, while the other may be called the reference sample. It is important, particularly when the dye swap technique is used, but also in general, to keep track of whether the target sample is labeled with the Cyanine 5 (Cy5, Red) dye or the Cyanine 3 (Cy3, Green) dye. In ***GESS***, this is done with a target sample dye variable. A column is entered into the spreadsheet containing the dye (Cy5 or Cy3) of the target sample. Only the values Cy5 or Cy3 may be entered into this column. This variable is required to run the Generic Two-Channel File Pre-Processing Engine procedure.

Output File Names Variable

When the Generic Two-Channel File Pre-Processing Engine is run, a new set of files may be generated for use in statistical analyses. The path and name for these newly created files may be entered into the output file names variable or an empty column may be specified. The files may be in different folders. This variable is required to obtain output for statistical analysis.

Blocks to be Deleted Variable

When two-channel pre-processing indicates there are some blocks (pin/print-tip groups) of some arrays that are of such poor quality that they need be removed, this may be done by listing those blocks in a column of the spreadsheet. Lists of the pin groups to be removed are entered into each cell, which corresponds to an output file, separated by spaces or commas.

For example, if it is desired that values from blocks 12, 21, and 33 be filtered (removed) from the output file of Row 7, then *12 21 33* may be entered in the cell of Row 7 of the Blocks to be Deleted Variable column.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options specify the variables that will be used in the analysis.

Two-Channel Input Files Specifications

These options are used specify the input files that are to be pre-processed.

Two-Channel File Name Variable

Select the variable that contains the list of the array files of the experiment.

The names and pathways of the files should appear in a column below this variable name on the spreadsheet.

Target Sample Dye Variable

Select the variable that identifies whether the target sample is in the Cyanine 5 (Cy5, red) channel or the Cyanine 3 (Cy3, green) channel.

This required variable must contain either Cy5 or Cy3 in each cell.

The reference sample is automatically assumed to be in the other channel. That is, if the target sample is in the Cyanine 5 (Cy5, red) channel, the reference sample is assumed to be in the Cyanine 3 (Cy3, green) channel, and vice versa.

In some experiments, all target samples are in the Cyanine 5 (Cy5, red) channel. In other experiments, all reference samples are in the Cyanine 5 (Cy5, red) channel. In dye-swap experiments, the target and reference samples are mixed among channels.

In paired sample experiments, the reference and target samples may refer instead to before and after treatment.

Pixel Summary Statistic Settings

These options determine the summary value that will be produced in the output files.

Foreground Statistic

Specify whether the Median, Mean, or Standard Deviation of the pixel intensities is to be used.

For example, if Median is selected here, the columns specified under Red (Cy5) F Median and Green (Cy3) F Median will be used.

Background Statistic

Specify whether the Median, Mean, or Standard Deviation of the pixel intensities is to be used.

For example, if Median is selected here, the columns specified under Red (Cy5) B Median and Green (Cy3) B Median will be used.

Create Target and Reference Using

Specify how the target and reference samples will be summarized.

For example, if 'log2(Foreground - Background)' is specified here, the target samples will be summarized by taking the target foreground statistic, subtracting the target background statistic, followed by taking the logarithm, base2. A similar calculation will occur for the reference.

RECOMMENDATION: We recommend log2(Foreground).

Expression Measure that is Output

The formula selected here indicates the formula that will be used to summarize the expression at each spot of the array. These values are output into a file that can be used in the statistical analyses procedures.

- **Target**

  Only the target sample summary is output. The reference sample is ignored.

- **Reference**

  Only the reference sample summary is output. The target sample is ignored.

- **Target – Reference**

  The reference sample summary is subtracted from the background sample summary.

- **Target – Reference – LOESS(Array)**

  The loess value based on the entire array is subtracted.

- **Target – Reference – LOESS(Block)**

  The loess value based on the current block is subtracted.

RECOMMENDATION: We recommend Target - Reference - LOESS(Block or Array).

Example: Suppose the target is in the Cy5 channel, 'Median' is selected under 'Foreground Statistic', 'log2(Foreground)' is selected under 'Create Target and Reference Using:', and 'Target - Reference' is selected under 'Expression Measure that is Output:'.

The output file will contain values using the formula log2(Cy5 F Median) - log2(Cy3 F Median), where F is for Foreground.

GES Output Files Specifications

These options are used to determine the location and naming of the output .ges files.

Folder in which Output Files will be Stored

Enter the path and name of the folder in which the newly created .ges files will be stored. The path may be typed directly or the Browse button may be used to locate the desired folder. New files will be created only if a variable is entered under Output File Names Variable.

In the default path "%mydocs_NCSS%\DATA\GESS", the designator "%mydocs_NCSS%" represents the path to the ***GESS*** personal data folder (commonly *C:\...\[My] Documents\NCSS \NCSS 2007*).

Output File Names Variable

Select the variable in which the path and names of the newly created .ges files will be entered for future statistical analyses. The path and folder of these .ges files is entered under Folder in which Output Files will be Stored.

A new file (to be used for statistical analyses) will be created for each row when the procedure is run.

If this variable is left blank, no new .ges files will be created.

Append to File Names

A sequence of characters may be entered here to append to the name of the created file.

For example, if the file has the name 'Slide1_10hours.txt' and 'log' is entered here, the newly created .ges file will be 'Slide1_10hours log.ges'.

If nothing is entered here, the file name will be the same as the name of the original file, but the extension will be replaced with '.ges'.

Overwrite existing output (.ges) files with new output (.ges) files

Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file.

For example, if Slide1.ges has already been created and a new Slide1.ges file is to be written, the new file will be Slide1 (2).ges if the Overwrite box is not checked.

File Specs Tab

The options on this panel control the labels used for scatter plots, spatial anomaly plots, and box plots.

Specify the Two-Channel File Column Heading Names

Column Names

A two-channel file may be made up of many columns. Each column may have a corresponding heading name. This program uses the heading name specified here to search for the appropriate column to read from the array file.

All column names necessary for pre-processing must be entered before pre-processing will run.

Column numbers (rather than names) may be entered here if 'None (Use Column Numbers)' is entered under 'Column Names are on Row Number'.

File Specifications

These specifications control where the column names are read, when to begin reading the data, and the delimiter used in the files.

Column Names are on Row Number

Enter the row of the array files on which the column names are found. This row must be the same on all the files.

If 'None (Use Column Numbers)' is entered here, then column numbers, rather than column names should be entered as individual column names.

Data begins on Row Number

Enter the row for the first line of gene expression data in the array files.

Delimiter

Enter the delimiter that separates the column names and expression values in the array files.

Reports Tab

The options on this panel control which reports and plots are generated.

Summary Reports

These options are used to determine the reports and report format that are output.

Specification Summary

Check this box to obtain a summary of the formula that is output to the output files, the subsets used, and the filters used.

Array Detail Summary

Check this box to obtain a row by row summary of names of files, numbers of filtered spots, array means and standard deviations, and deleted blocks.

Mean Decimals

Specify the number of decimals used for means in the numeric portion of the Array Detail Summary. The number of decimal places is used for display only. It does not change the internal precision of the data.

SD Decimals

Specify the number of decimals used for standard deviations in the numeric portion of the Array Detail Summary. The number of decimal places is used for display only. It does not change the internal precision of the data.

Select Plots

The following options are used to determine which plots will be displayed.

Spatial Anomaly Plot

Check this box to indicate that you want this whole array spatial anomaly plot displayed for all arrays. The spatial anomaly plot gives a spatial view of the entire array for the corresponding measurement. Intensities are separated into four color groupings that reflect four percentile groups. The settings of the spatial anomaly plot are specified under the Spatial Plot tab.

Box Plot - Arrays

Check this box to indicate that you want to display side-by-side box plots comparing all arrays for this measurement. The settings of the box plot comparing array measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Box Plot - Blocks

Check this box to indicate that you want to display side-by-side box plots comparing blocks (pin/print-tip groups) for this measurement for all arrays. The settings of the box plot comparing blocks (pin/print-tip groups) measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Box Plot - Subsets

Check this box to indicate that you want to display side-by-side box plots comparing subsets (control groups) to the primary group of spots for this measurement for all arrays. The settings of the box plot comparing subsets (control groups) measurements are specified under the Box Plot, Min-Max, and Labels tabs.

Scatter Plot – M vs A

Check this box to indicate that you want to display the M vs A plot for each array.

Sometimes called the Ratio-Intensity (R-I) plot, this plot is used to monitor dye bias. The Y axis specifies the value of M, the difference in the intensity summaries of the two samples, Target Summary - Reference Summary. M is a pneumonic for 'minus'. The X axis indicates the value of A, the average intensity summary of the two samples, (Target + Reference)/2. A is a pneumonic for 'add, or average, or abundance'. The loess line is a moving weighted regression average.

Scatter Plot – M’ vs A

Check this box to indicate that you want to display the M' vs A plot for each array.

This plot may be compared to the M vs A plot to see the effect of whole array loess subtraction on dye bias. The Y axis specifies the value of M', where M' = M - whole array loess value. The X axis indicates the value of A, which is the same as in the M vs A plot. The new loess line of the M' values is included.

Scatter Plot – M’’ vs A

Check this box to indicate that you want to display the M'' vs A plot for each array.

This plot may be compared to the M vs A plot to see the effect of block loess subtraction on dye bias. The Y axis specifies the value of M'', where M'' = M - block loess value. The X axis indicates the value of A, which is the same as in the M vs A plot. The new loess line of the M'' values is included.

Scatter Plot – T vs R

Check this box to indicate that you want to display the Target vs Reference plot for each array.

This plot is similar to the M vs A Plot. The Y axis shows the intensity summary of the Target sample. The X axis indicates the intensity summary of the Reference Sample.

Filters 1 Tab

The options on this panel control the weak signal filters that will be used.

Filter

Check this box to activate this filter. Spots meeting the criterion will be omitted from the output dataset, but will still be included in pre-processing graphical displays.

Filter Boundary

Specify the filter boundary value. When the spot value is below this boundary value, the output for this spot is omitted from the output dataset. All spots will still be included in pre-processing graphical displays.

Filters 2 Tab

The options on this panel control the saturation, standard deviation, pin group, and negative filters that will be used.

SD Filter

Check this box to activate this filter. Spots meeting the criterion will be omitted from the output dataset, but will still be included in pre-processing graphical displays.

SD Filter Boundary

Specify the filter boundary value. When the spot value is above this boundary value, the output for this spot is omitted from the output dataset. All spots will still be included in pre-processing graphical displays.

Blocks to be Deleted Variable

Specify the name of the variable that contains a list of blocks (pin/print-tip groups) to be deleted from the output dataset. This variable should have been created on the spreadsheet. Blocks (pin/print-tip groups) are excluded within arrays according to this variable, not across arrays. Blocks (pin/print-tip groups) should be separated by spaces as entries in the cells of the column under this variable name.

This variable is usually created after an initial run to identify problematic blocks (pin/print-tip groups).

Subsets 1 - 9 Tabs

The options on this panel control the names and lists of subsets.

Subset (1 – 9) Name

The name of the gene (spot) subset is entered here.

Plots comparing subsets can be obtained by checking the boxes next to 'Box Plot - Subsets' under the Reports tab. The name chosen here will appear on these plots.

EXAMPLE: To determine whether or not the microarray is functioning properly, it is common to introduce spike-in control DNA into the sample. Spike-in control DNA has a known relative intensity, e.g., 5-fold (target 5 times the reference sample), and corresponds to carefully chosen spots on the array. A list of Spike-in Controls could be given here. Values for the spike-in controls may be compared using box plots to negative controls, positive controls, blank spots, etc. to show that, in fact, the spike-in controls have higher relative intensity values. This would indicate a properly functioning microarray.

Spots in this Subset

Enter a list of genes (spots) that are to be in this subset. The genes (spots) may be entered directly, or the * character may be used to specify all genes with a particular beginning. The gene names or IDs entered in this list must be in the column specified in Spot Name From box on the Variables tab.

EXAMPLES:

Blank

spike1

spike3

spike5

spike* (all names beginning with spike)

AA44719

NM_00582

NM_04762

NM_27564

cntrl* (all names beginning with cntrl)

file(C:\Microarray\genelist.txt) (all names in the genelist.txt file)

var(OutputGenes) (all names in the spreadsheet variable with the variable name OutputGenes)

Output for Analysis

If this box is checked, the spots in this subset will be included in the output file for future statistical analyses. If this box is not checked, these spots will be removed from the output file.

(Plotting) Symbol

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the plotting symbol. This plotting symbol will be used in the selected array quality graphics.

Min-Max Tab

The options on this panel control the minimum and maximum values for the axes of the box plots and scatter plots.

Axis Minimum

Specify the value to be displayed as the minimum on this axis. Data values less than this amount will be ignored. If this value is left blank, the minimum will be determined from the data.

Axis Maximum

Specify the value to be displayed as the maximum on this axis. Data values greater than this amount will be ignored. If this value is left blank, the maximum will be determined from the data.

Labels Tab

The options on this panel control the labels used for scatter plots, spatial anomaly plots, and box plots.

Short Label

Enter here the text that is to be used for the short labels of box plots, spatial anomaly plots, and scatter plots.

Long Label

Enter here the text that is to be used for the long labels of box plots, spatial anomaly plots, and scatter plots. The default is based on the entries under Pixel Summary Statistic Settings of the Variables Tab.

Spatial Plot Tab

The options on this panel control the features of the spatial anomaly plot.

Heat Map Settings

These settings are used to control the appearance of the heat map and its legend.

Heat Map Colors and Scale

Click on the heat map color bar or the button to the right to change the colors and/or scale of the heat map.

Label

Enter text here for the legend label.

Number of Values

This is the number of reference values printed along the right side of the heat map legend.

Show Legend

Specify whether to show the legend.

Value Format

This option specifies the characteristics of the reference numbers shown next to the heat map legend.

It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot.

It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Plot Settings

These options are used to specify the appearance of the heat map surroundings.

Plot Style file

A plot style file sets all plot options that are not set directly by this procedure.

Interior Color

Specify the interior color of the spatial anomaly plot.

Background Color

Specify the background color of the spatial anomaly plot.

Plot Titles

Enter text here for the designated title or label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by the long version of the selected intensity summary.

{Y} is replaced by the short version of the selected intensity summary.

{Z} is replaced by the appropriate array number.

Box Plot Tab

The options on this panel control attributes of the box plots.

Vertical and Horizontal Axes

These options control the vertical and horizontal axes attributes.

Axis Labels

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by the appropriate values of the corresponding X axis grouping variable.

{Y} is replaced by the short version of the Y axis intensity summary.

Ref. Number Format

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Major Ticks

Specify the number of large tickmarks and optional grid lines along this axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along this axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along this axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the axis. Thus, checking the Y Grid Lines will actually cause horizontal grid lines to appear.

Box Plot Settings

These options are used to specify the appearance of the box plots.

Box Plot Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created in the Box Plots procedure.

Box Percent Space

When the Box Width (or Bar Width) option is set to Percent Space in the Box Plot Style File selected, this value specifies the percent of the length of the axis that is empty space instead of bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**

  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**

  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Interior Color

Specify the interior color of the plot.

Background Color

Specify the background color of the plot.

Titles

Enter text for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{G} is replaced by the long version of the Y axis intensity summary.

{Y} is replaced by the short version of the Y axis intensity summary.

{Z} is replace by the appropriate array number.

Box Plot Colors

The options are used to specify the colors of the box plots.

Fill Color

The color used to fill this object. Click to change.

Outline (Border) Color

The color used to outline the object. Click to change.

Line Color

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Scatter Plot Tab

The options on this panel control the main features of the M vs A and T vs R plots.

Vertical and Horizontal Axes

These options control the vertical and horizontal axes attributes.

Ref. Number Format

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Major Ticks

Specify the number of large tickmarks and optional grid lines along this axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along this axis.

Grid Lines

Check this option to display grid lines at the major tickmarks along this axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the axis. Thus, checking the Horizontal Axis Grid Lines will actually cause vertical grid lines to appear.

Scatter Plot Settings

These options are used to specify the appearance of the scatter plots.

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. Scatter plot style files are created in the Scatter Plots procedure.

Symbol

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the points plotted on the scatter plot.

Diamond

Check this box to display a diamond showing the physical boundaries of M on the M vs A, M' vs A, and M'' vs A scatter plots.

Zero

Check this box to display a horizontal line at 0 on the M vs A, M' vs A, and M'' vs A scatter plots.

45 Degree

Check this box to display a 45 degree line on the T vs R scatter plot only. The 45 degree line shows where T and R are equivalent.

Interior Color

Specify the interior color of the plot.

Background Color

Specify the background color of the plot.

Plot Titles

Enter text here for the designated title or label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{M} is replaced by the long version of the X axis intensity summary.

{S} is replaced by the long version of the Y axis intensity summary.

{X} is replaced by the short version of the X axis intensity summary.

{Y} is replaced by the short version of the Y axis intensity summary.

{Z} is replaced by the appropriate array number.

Loess Options

These options are used to determine whether a loess line is included, its appearance, and the details of how it is computed.

Include Loess Curve

Check this option to display a Loess smooth line.

The locally-weighted, robust regression (loess) smooth is a popular, computer-intensive technique that usually provides a reasonable smoothing of your data without being overly sensitive to outliers. A reasonable smooth is one that travels more or less through the middle of the data. The degree of smoothing is controlled by the Loess % N option.

Loess Order

The order of the polynomial fit in the Loess procedure. Select '1' for a linear fit or '2' for a quadratic fit.

RECOMMENDED: 2 - Quadratic

Loess % N

The percent of the dataset to be used at each Loess calculation.

RECOMMENDED: 40

RANGE: 1 to 99

Number of Points

Specify the number of points at which the Loess line is evaluated. This affects the granularity of the lines. More points imply smoother lines. The number of points selected here may considerably affect the run time.

RANGE: 20 to 2000.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Pre-Processing Generic Two-Channel Files

This section presents an example of how to pre-process five two-channel files, without involving subsets or filters.

Each of the files represents a summary file that is output from image processing software. Each of the files has 23 columns with the following headings. The column headings are located on the first line (i.e., there is no header) of each file and are separated by tabs.

Pin
Column
Row
Name
X location
Y location
Diameter
Cy5ForeMedian
Cy5ForeMean
Cy5ForeSD
Cy5BackMedian
Cy5BackMean
Cy5BackSD
Cy3ForeMedian
Cy3ForeMean
Cy3ForeSD
Cy3BackMedian
Cy3BackMean
Cy3BackSD
%UsedFore
%UsedBack
ForePixels
NumPixels

The data for each of these columns begins on line two. There are 8 pin groups and 3,360 genes. The spreadsheet data used are recorded in the TC_Ex1 dataset.

To run this example, take the following steps or load the **Example 1** template from the Generic Two-Channel File Pre-Processing Engine Template tab.

1 Open the TC_Ex1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **TC_Ex1.S0**.
- Click **Open**.

2 Open the Generic Two-Channel File Pre-Processing Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Generic Two-Channel Files**. The Generic Two-Channel File Pre-Processing Engine procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Generic Two-Channel File Pre-Processing Engine window, select the **Variables** tab.
- Set the **Two-Channel File Name Variable** to **InputFile**.
- Set the **Target Sample Dye Variable** to **Target**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **OutputFile**.
- Leave **Append to File Names** blank.

4 Specify the Pixel Summary Statistic Settings.

- Continuing on the Variables tab, set **Create Target and Reference Using** to **log2(Foreground)**.
- Set **Expression Measure that is Output** to **Target – Reference – LOESS(Block)**.
- Set the **Foreground Statistic** to **Median**.

5 Enter the Column Names.

- Select the **File Specs tab**.
- Enter the following names:
  Red (Cy5) F Median: Cy5ForeMedian
  Red (Cy5) F Mean: Cy5ForeMean
  Red (Cy5) F SD: Cy5ForeSD
  Red (Cy5) B Median: Cy5BackMedian
  Red (Cy5) B Mean: Cy5BackMean
  Red (Cy5) B SD: Cy5BackSD
  Green (Cy3) F Median: Cy3ForeMedian
  Green (Cy3) F Mean: Cy3ForeMean
  Green (Cy3) F SD: Cy3ForeSD
  Green (Cy3) B Median: Cy3BackMedian

Green (Cy3) B Mean: Cy3BackMean
Green (Cy3) B SD: Cy3BackSD
Gene Name: Name
Block (Pin Group): Pin
X Coordinate: X location
Y Coordinate: Y location

- Under **Column Names are on Row Number**, enter **1**.
- Under **Data begins on Row Number**, enter **2**.
- Under **Delimiter**, select **Tab**.

6 Specify the Reports.

- Select the **Reports tab**.
- Check the boxes next to **Specification Summary** and **Array Detail Summary**.
- Check the **Cy5** and **Cy3** boxes next to **Spatial Anomaly Plot**, **Box Plot – Arrays**, and **Box Plot – Blocks**.
- Check the **M** and **M’’** boxes to the right of **Scatter Plot – vs A**.

7 Specify the Spatial Anomaly Plot Settings.

- Select the **Spatial Plot tab**.
- Change the **four group symbols** to **solid circles** with **radius 40**.

8 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Expression Formula Output for Analysis

```
Expression Formula Output for Analysis

Log2(TFMd)-Log2(RFMd)-Loess(Block)

where

Log2: Logarithm Base 2
TF: Target Sample, Foreground Region of Spot
RF: Reference Sample, Foreground Region of Spot
Md: Median Pixel Intensity
Loess(Block): Loess Values Calculated Within Each Block
```

This report displays the formula that is used when the output files are created. In this case, the formula may be read as ‘log base 2 of the target foreground median minus log base 2 of the reference foreground median minus the within block loess value.

Value for Spot Was Deleted (Filtered) If

```
Value for Spot Was Deleted (Filtered) If

No Filters Selected
```

This report shows that no filters were selected.

Subset Summary

Subset	Subset Values Output for Analysis?
No subsets were used.	

This report shows that no subsets were created.

Input File Summary

Input File Summary

Row	Input File
1	…\Data\GESS\TC\TC1.txt
2	…\Data\GESS\TC\TC2.txt
3	…\Data\GESS\TC\TC3.txt
4	…\Data\GESS\TC\TC4.txt
5	…\Data\GESS\TC\TC5.txt

This report shows a list of the input file paths.

Output File Summary

Output File Summary

Row	Output File
1	…\data\gess\TC1.ges
2	…\data\gess\TC2.ges
3	…\data\gess\TC3.ges
4	…\data\gess\TC4.ges
5	…\data\gess\TC5.ges

This report shows a list of the output file paths. These are the names of the files that will be used as input for statistical analyses.

Numeric Array Summary - Foreground

Numeric Array Summary - Foreground

Row	Input File Name	Mean of Target Foreground Medians	Standard Deviation of Target Foreground Medians	Mean of Reference Foreground Medians	Standard Deviation of Reference Foreground Medians
1	TC1.txt	1202.6	921.9	1758.5	1439.4
2	TC2.txt	612.6	516.4	923.5	1065.5
3	TC3.txt	714.9	723.1	1084.6	1458.6
4	TC4.txt	800.9	691.4	1029.1	1118.0
5	TC5.txt	907.1	1122.9	1067.5	1497.5

Note: Means and standard deviations summarize all spots on the array, before filtering.

This report shows the whole array foreground region means and standard deviations for target and reference samples.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Mean of Target Foreground Medians

For the target sample, this is the average of all median pixel intensities of the foreground regions of the entire array.

Standard Deviation of Target Foreground Medians

For the target sample, this is the standard deviation of all median pixel intensities of the foreground regions of the entire array.

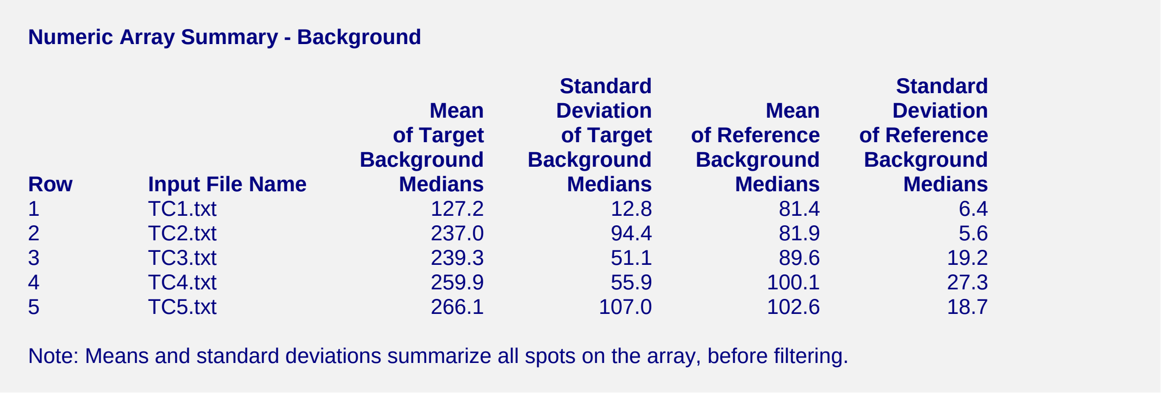
Mean of Reference Foreground Medians

For the reference sample, this is the average of all median pixel intensities of the foreground regions of the entire array.

Standard Deviation of Reference Foreground Medians

For the reference sample, this is the standard deviation of all median pixel intensities of the foreground regions of the entire array.

Numeric Array Summary - Background



```
Numeric Array Summary - Background

                                                 Standard                      Standard
                                    Mean        Deviation         Mean        Deviation
                               of Target        of Target  of Reference    of Reference
                              Background       Background    Background      Background
Row      Input File Name         Medians          Medians       Medians         Medians
1        TC1.txt                   127.2             12.8          81.4             6.4
2        TC2.txt                   237.0             94.4          81.9             5.6
3        TC3.txt                   239.3             51.1          89.6            19.2
4        TC4.txt                   259.9             55.9         100.1            27.3
5        TC5.txt                   266.1            107.0         102.6            18.7

Note: Means and standard deviations summarize all spots on the array, before filtering.
```

This report shows the whole array background region means and standard deviations for target and reference samples.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Mean of Target Background Medians

For the target sample, this is the average of all median pixel intensities of the background regions of the entire array.

Standard Deviation of Target Background Medians

For the target sample, this is the standard deviation of all median pixel intensities of the background regions of the entire array.

Mean of Reference Background Medians

For the reference sample, this is the average of all median pixel intensities of the background regions of the entire array.

Standard Deviation of Reference Background Medians

For the reference sample, this is the standard deviation of all median pixel intensities of the background regions of the entire array.

Spot Summary

```
Spot Summary
                                                          Total
                                        Total                Filtered      Total
                                        Filtered   Missing       and      Active      Total
Row           Input File Name           Spots      Values    Missing     Spots      Spots
1             TC1.txt                       0          0          0       3360       3360
2             TC2.txt                       0          0          0       3360       3360
3             TC3.txt                       0          0          0       3360       3360
4             TC4.txt                       0          0          0       3360       3360
5             TC5.txt                       0          0          0       3360       3360
```

This report shows a summary of filtered spots, missing values, active and total spots.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Total Filtered Spots

This is number of spots that were filtered. The specifics of the filter are found in the next summary.

Missing Values

This is number of missing values among all spots.

Total Filtered and Missing

This is the sum of the Total Filtered Spots and the Missing Values.

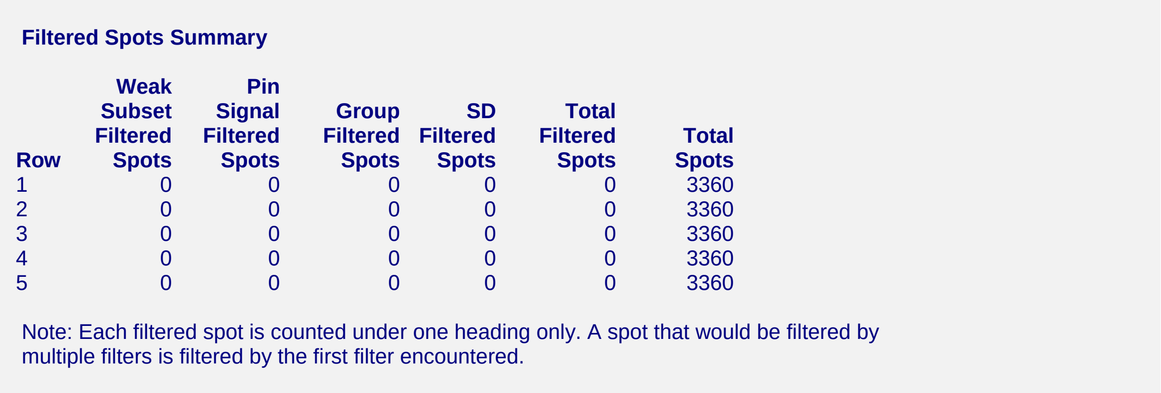
Total Active Spots

This is the number of spots that are not filtered, nor missing.

Total Spots

This is the total number of spots on the array.

Filtered Spots Summary



Filtered Spots Summary

	Weak	Pin				
	Subset	Signal	Group	SD	Total	
	Filtered	Filtered	Filtered	Filtered	Filtered	Total
Row	Spots	Spots	Spots	Spots	Spots	Spots
1	0	0	0	0	0	3360
2	0	0	0	0	0	3360
3	0	0	0	0	0	3360
4	0	0	0	0	0	3360
5	0	0	0	0	0	3360

Note: Each filtered spot is counted under one heading only. A spot that would be filtered by multiple filters is filtered by the first filter encountered.

This report shows a detailed summary of all filtered spots.

Row

This is the row of the array in the spreadsheet.

Input File Name

This is the name of the file without the path.

Subset Filtered Spots

This is the total number of spots that were filtered because they were members of a deleted subset.

Weak Signal Filtered Spots

This is the total number of spots that were filtered based on one or more of the twelve weak signal filters.

Saturation Filtered Spots

This is the total number of spots that were filtered based on one or more of the two saturation filters.

SD Filtered Spots

This is the total number of spots that were filtered based on one or more of the four standard deviation filters.

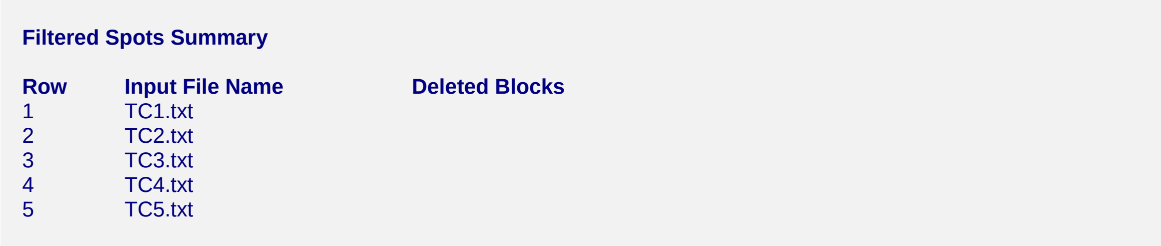
Total Filtered Spots

This is number of spots that were filtered.

Total Spots

This is the total number of spots on the array.

Deleted Blocks Summary



Filtered Spots Summary

Row	Input File Name	Deleted Blocks
1	TC1.txt	
2	TC2.txt	
3	TC3.txt	
4	TC4.txt	
5	TC5.txt	

This report shows a summary of pin groups that were filtered within each array.

Row

This is the row of the array in the spreadsheet.

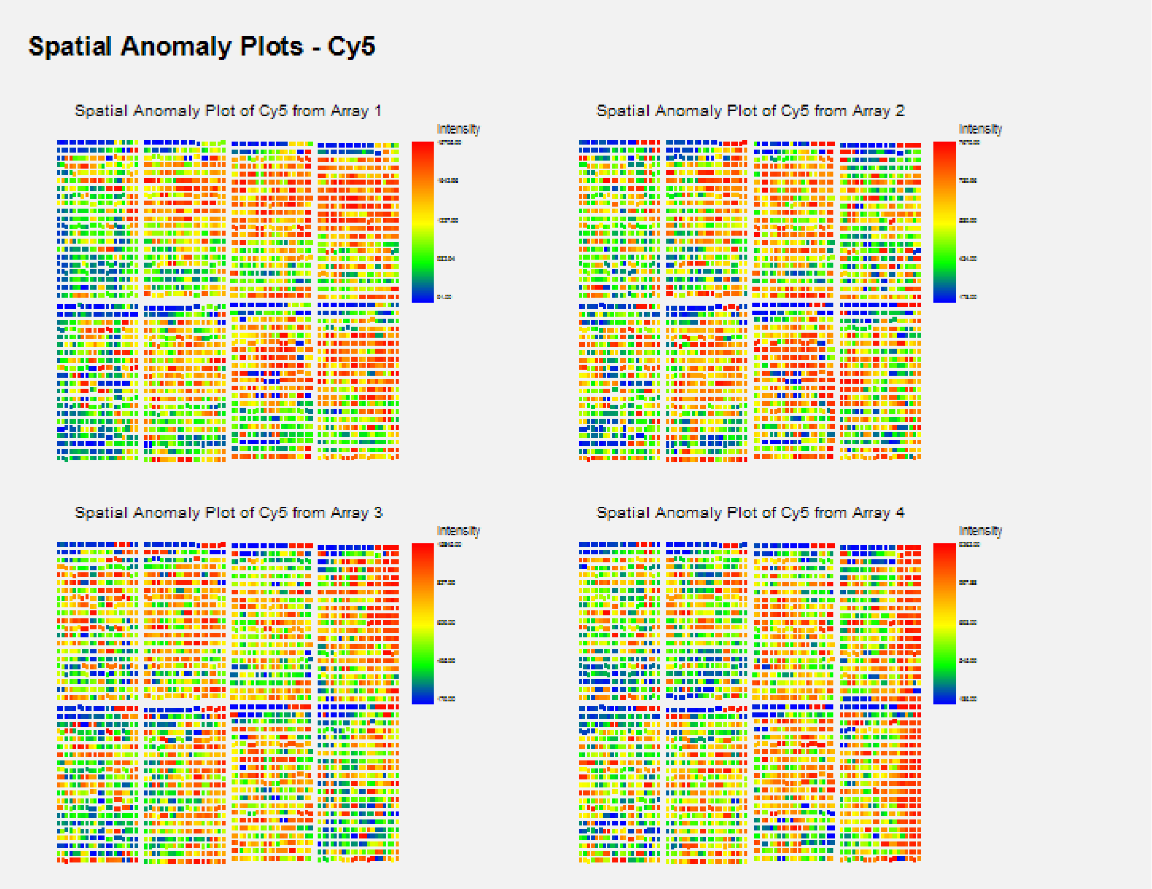
Input File Name

This is the name of the file without the path.

Deleted Blocks

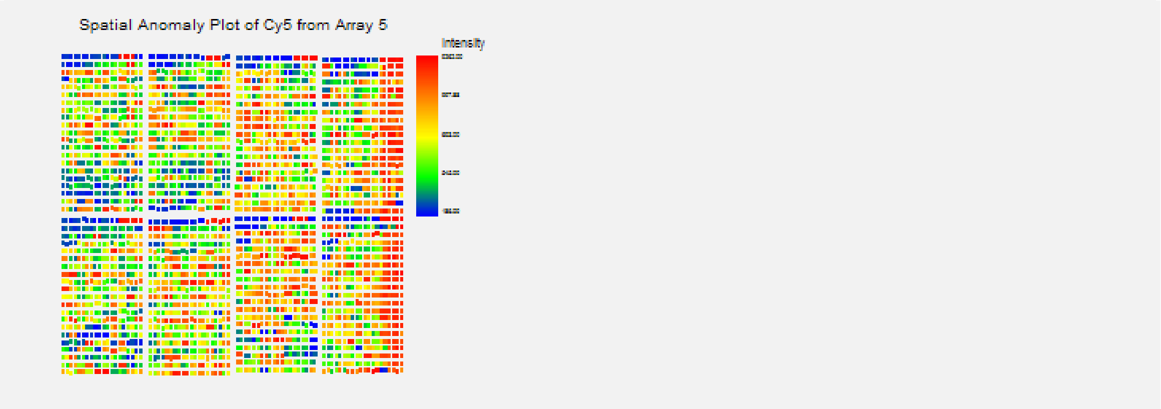
These are the blocks for which the all spots are deleted.

Spatial Anomaly Plots – Cy5



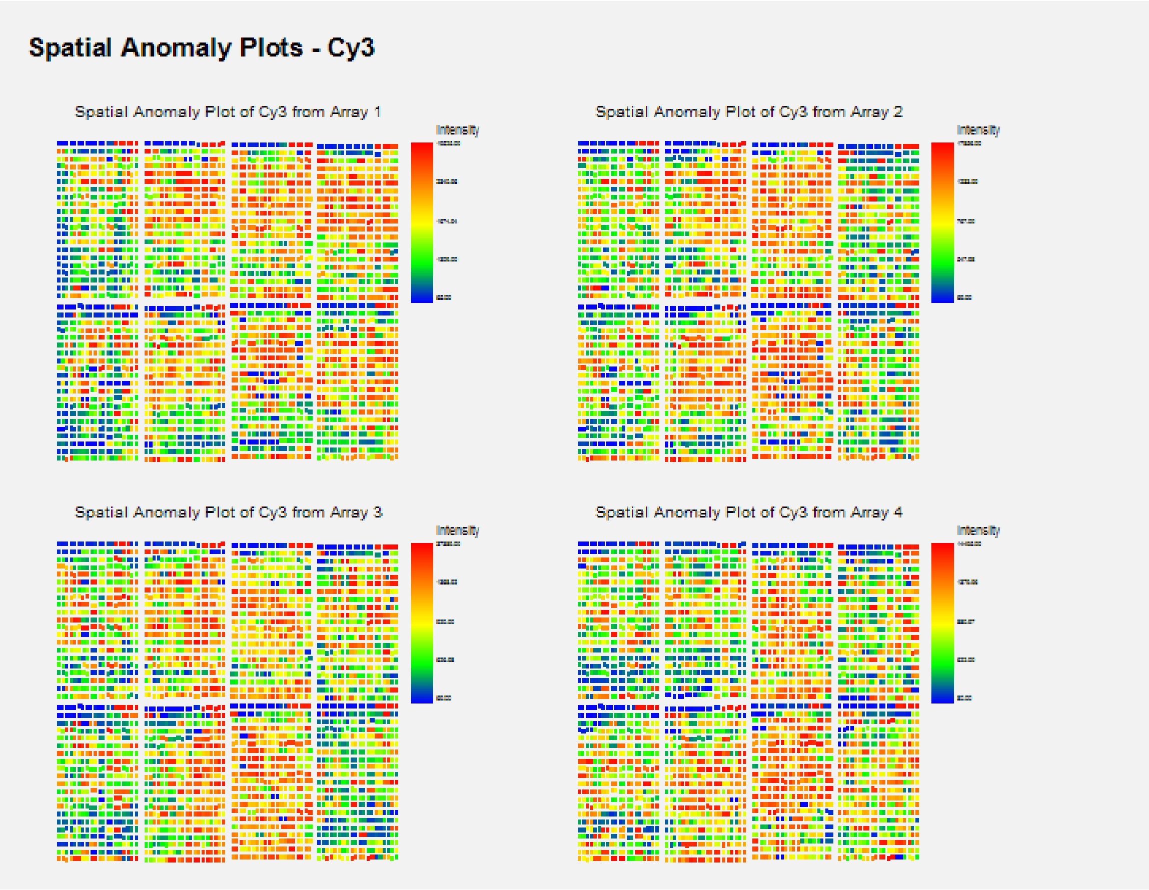
Spatial Anomaly Plots - Cy5

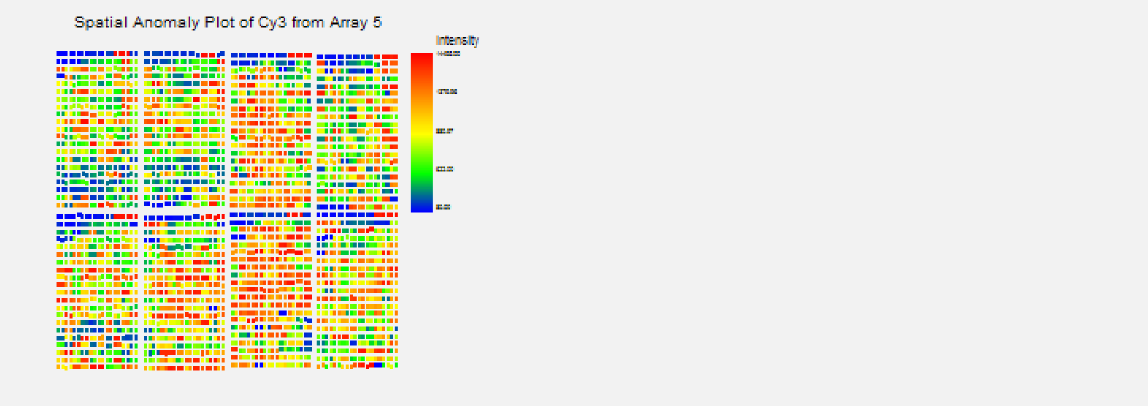
Spatial Anomaly Plot of Cy5 from Array 1
Intensity
Spatial Anomaly Plot of Cy5 from Array 2
Intensity
Spatial Anomaly Plot of Cy5 from Array 3
Intensity
Spatial Anomaly Plot of Cy5 from Array 4
Intensity



This report shows a spatial representation of the Cyanine 5 (Cy5, Red) median foreground intensities. There appear to high expression arrays on the right side of arrays 4 and 5. There is also a highly expressed region in array 1.

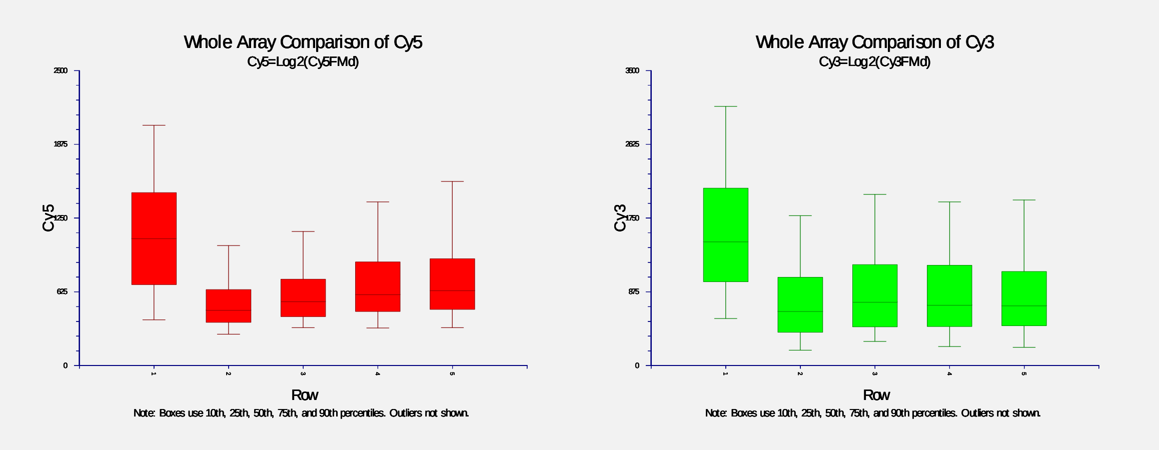
Spatial Anomaly Plots – Cy3





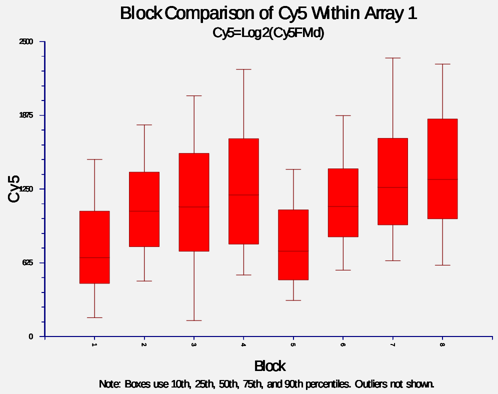
This report shows a spatial representation of the Cyanine (Cy3, Green) median foreground intensities. There are several patterns to indicate spatial problems.

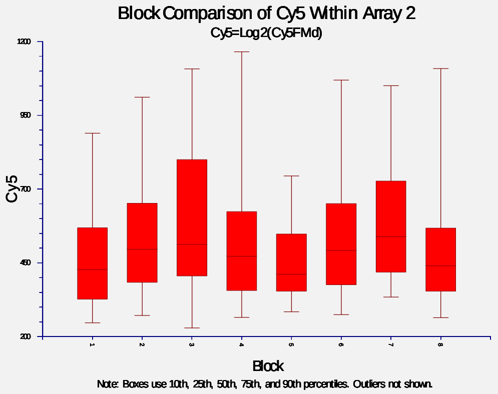
Array Comparison Section

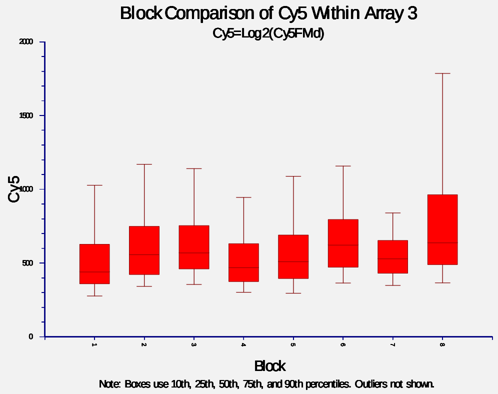


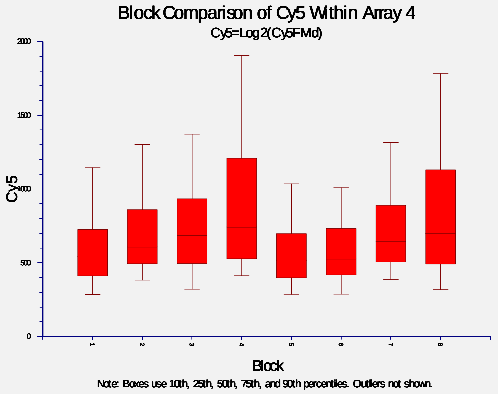
These plots allow comparison of Log2(Foreground Median) of the 5 arrays for both Cyanine 5 (Cy5, Red) and Cyanine 3 (Cy3, Green) channels. Array 1 seems to have an unusually high expression pattern.

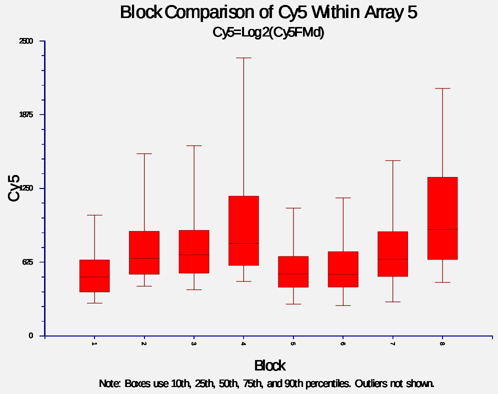
Block Comparison Section – Cy5





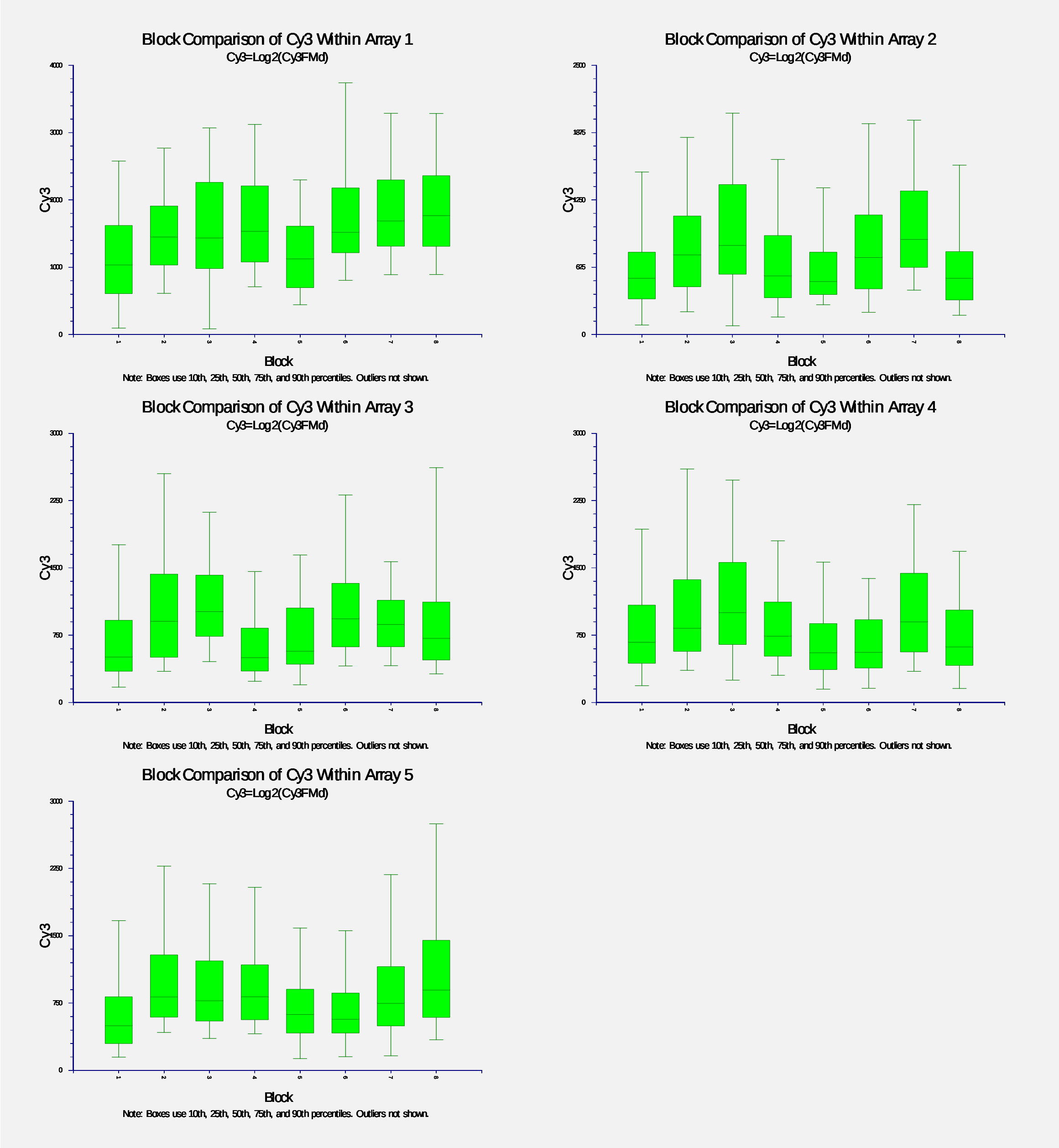






These plots allow comparison of Log2(Foreground Median) across all blocks within each array for the Cyanine 5 (Cy5, Ged) channel. There is an upward trend in expression from left blocks to right blocks.

Block Comparison Section – Cy3



Block Comparison of Cy3 Within Array 1

Cy3=Log2(Cy3FMd)

Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Block Comparison of Cy3 Within Array 2

Cy3=Log2(Cy3FMd)

Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Block Comparison of Cy3 Within Array 3

Cy3=Log2(Cy3FMd)

Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

Block Comparison of Cy3 Within Array 4

Cy3=Log2(Cy3FMd)

Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

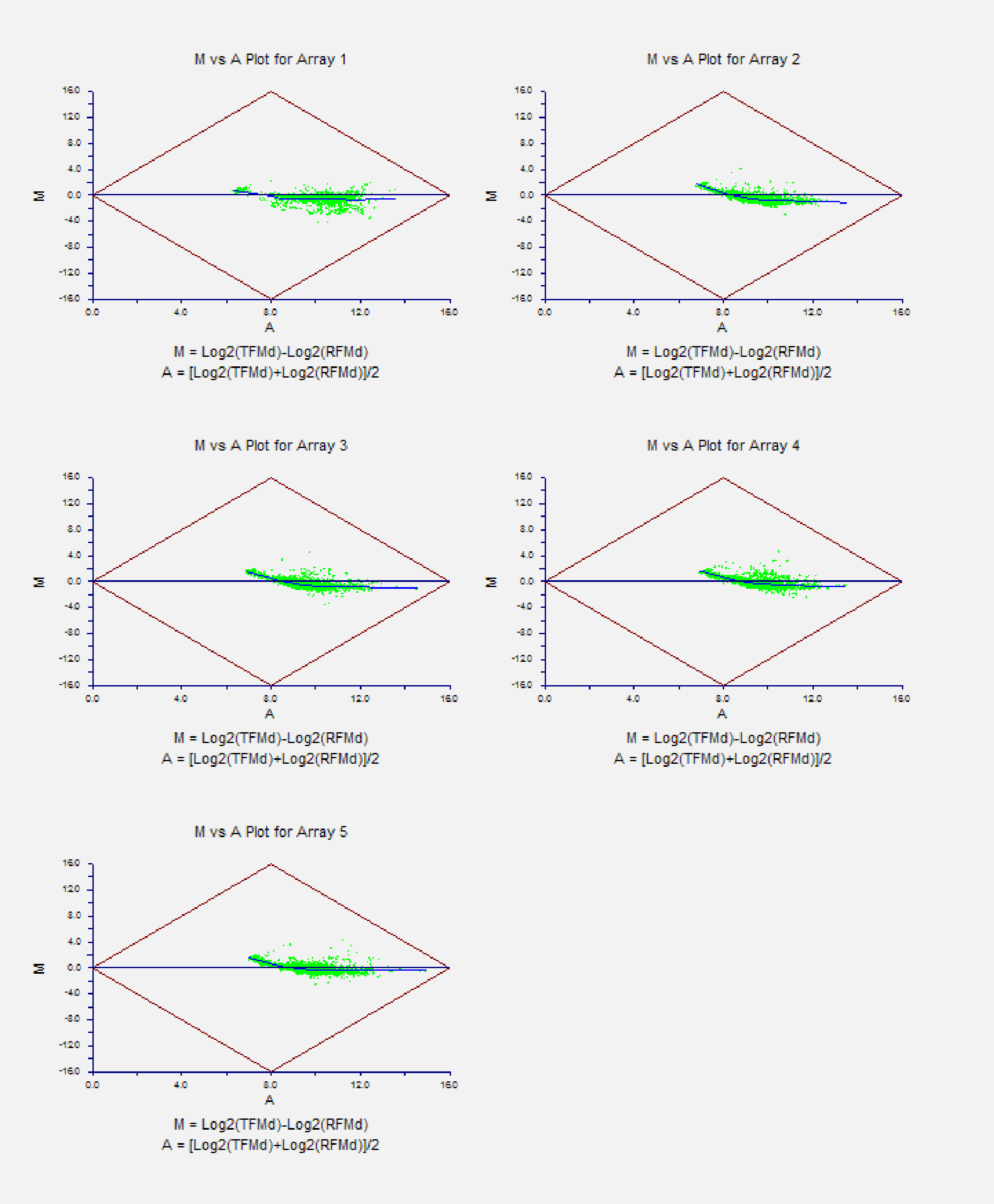
Block Comparison of Cy3 Within Array 5

Cy3=Log2(Cy3FMd)

Note: Boxes use 10th, 25th, 50th, 75th, and 90th percentiles. Outliers not shown.

These plots allow comparison of Log2(Foreground Median) across all blocks within each array for the Cyanine 3 (Cy3, Green) channel. The upward trend is not as pronounced in the Cy3 channel.

M vs A Section



M vs A Plot for Array 1

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 2

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 3

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 4

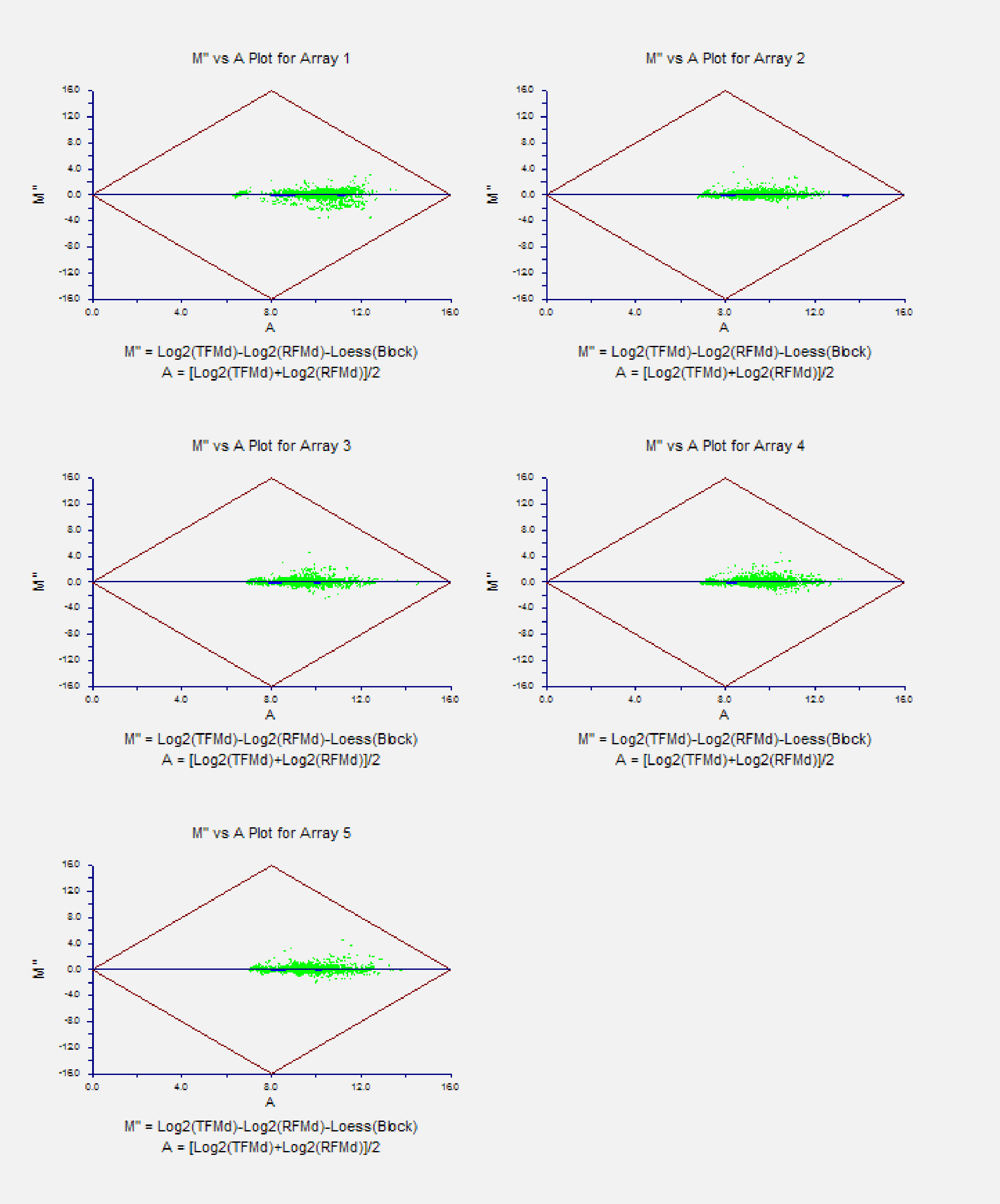
M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

M vs A Plot for Array 5

M = Log2(TFMd)-Log2(RFMd)
A = [Log2(TFMd)+Log2(RFMd)]/2

These M vs A plots can be used to monitor dye bias. The M value (Y-axis) is the Target minus Reference difference in Log2(Foreground Median) intensities. The A value (X-axis) is the average of the Log2(Foreground Median) intensities. The diamond shows the physical limits of the M vs A coordinates when a 16-bit scanner is used. The loess line is overlaid. If there is no dye bias, the loess line will be near zero.

M’’ vs A Section



These M’’ vs A plots show the dye bias correction obtained when subtracting the within block loess value. The loess line is overlaid, but is difficult to see since it is at or near the zero line, indicating the dye bias has been removed.

Chapter 131

Generic Expression Data Import Engine

Introduction

This chapter describes the process of importing expression data from generic, delimited text files using the ***GESS*** Generic Expression Data Import Engine. This procedure should be used to import expression data files generated by outside pre-processing software. Common expression data files that can be imported using this procedure have the extensions .txt, .dat, .csv, etc. Files with other extensions can be imported, but they must have a delimited text file format. Affymetrix .chp files should be imported using the Affymetrix CHP File Import Engine, not this import engine.

The engine works by taking each input file from the designated column on the spreadsheet and creating a single output (.ges) expression file for each array or column of expression data in the input file. Multiple files may be imported on a single run of the ***GESS*** Generic Expression Data Import Engine. These files may have various numbers of expression data columns per file, but all files must have the same general file format and structure. These format requirements are presented in detail in this chapter.

Input File Specifications

The ***GESS*** Generic Expression Data Import Engine was designed for maximum flexibility by allowing for a wide variety of input file formats. However, this procedure requires that all input files from a single importing run be of the same general format and file type. Below is a list of the file requirements for using this procedure:

1. All files must be delimited text files.
2. All files must have exactly the same format and content structure.
3. All files must be delimited by the same character (Tab, Comma, Space, or Semicolon).
4. A column of gene (probeset) names must be included in each file. This column of gene names (specified as the "Gene Names Column") must be in the same location for all input files and located to the left of the first column of gene expression data.

5. At least one column of expression values must be included in each file. Often, a single text file contains multiple columns of expression data separated by one or more columns containing other information such as standard deviations, presence/absence calls, etc. This other information is ignored by ***GESS*** during generic data importing. All files must contain the same number of ignored columns (if any) between expression data columns. The number of columns between each expression data column is specified as "Columns Skipped".
6. The first column of expression values (specified as the "First Data Column") must be located to the right of the column containing gene (probeset) names.
7. Column titles (if present) must be on the same row for all input files. The row containing column titles is specified as the "Column Names Row". The first row of data must immediately follow the column names row. The column titles are used as the basis for the output (.ges) file names. If column titles are absent in any input file, they must be absent in all input files. In the event that column titles are absent, ***GESS*** will create appropriate output (.ges) file names based on the input file names and data column numbers.

Compatible File Format Examples

Below are examples of compatible import files. All files are assumed to be delimited text files containing expression values.

Compatible Example 1: Tab-Delimited Text File without Column Titles

```
H-1     147.33
H-2     94.19
H-3     132.58
H-4     45.23
H-5     67.08
.       .
.       .
.       .
```

This file is the most basic file that can be imported into ***GESS***. The file has a single column of gene names followed by a single column of expression estimates. ***GESS*** will automatically generate an output (.ges) file name based on the name of this file. The output file will have the name, "[Input File Name].ges".

File Specifications

Delimiter	Tab
Columns Skipped	0
Column Names Row	None
Gene Names Column	1
First Data Column	Gene Names Column + 1

Compatible Example 2: Tab-Delimited Text File without Column Titles

```
Gene    Chip1.cel
H-1     147.33
H-2     94.19
H-3     132.58
H-4     45.23
H-5     67.08
.       .
.       .
.       .
```

This file is the same as Example 1, except that it has column titles. The output file will have the name, "Chip1.ges".

File Specifications

Delimiter	Tab
Columns Skipped	0
Column Names Row	1
Gene Names Column	1
First Data Column	Gene Names Column + 1

Compatible Example 3: Tab-Delimited Text File with Multiple Columns

```
Gene    Chip1.cel     Chip2.cel     Chip3.cel     Chip4.cel     Chip5.cel
H-1     178.93        7.27          38.75         80.60         18.21
H-2     199.85        118.82        58.76         97.01         191.03
H-3     157.68        145.29        34.53         33.88         190.44
H-4     45.38         181.77        48.37         99.40         21.00
H-5     146.90        68.98         184.92        197.25        105.40
.       .             .             .             .             .
.       .             .             .             .             .
.       .             .             .             .             .
```

This file type is generated by many pre-processing programs that read several intensity files at a single time. The file has multiple columns of expression data. The output files will have the names, "Chip1.ges", "Chip2.ges", "Chip3.ges", etc.

File Specifications

Delimiter	Tab
Columns Skipped	0
Column Names Row	1
Gene Names Column	1
First Data Column	Gene Names Column + 1

Compatible Example 4: Tab-Delimited Text File without Column Titles and with an Intervening Text Column

```
H-1    A      178.93       7.27         38.75        80.60        18.21
H-2    B      199.85       118.82       58.76        97.01        191.03
H-3    C      157.68       145.29       34.53        33.88        190.44
H-4    D      45.38        181.77       48.37        99.40        21.00
H-5    E      146.90       68.98        184.92       197.25       105.40
.      .      .            .            .            .
.      .      .            .            .            .
.      .      .            .            .            .
```

In this example, the first data column (column 1) does not directly follow the gene names column (column 3). The first data column is column 3. Without column titles, ***GESS*** will automatically generate output file names based on the name of this file and the column numbers. The output files will have the names, "[File Name]_Col2.ges", "[File Name]_Col3.ges", "[File Name]_Col4.ges", etc.

File Specifications

Delimiter	Tab
Columns Skipped	0
Column Names Row	None
Gene Names Column	1
First Data Column	3

Compatible Example 5: Tab-Delimited Text File with Header and Intervening Columns of Text

```
Date: June 6, 2006
CDF File: Test
Technician: Braden
Gene   Chip1.cel  Call   Chip2.cel   Call   Chip3.cel    Call
H-1    135.40     P      67.90       A      103.23       P
H-2    0.06       A      9.42        A      106.65       P
H-3    6.15       A      25.13       A      62.14        A
H-4    78.24      A      102.89      P      86.59        A
H-5    30.78      A      176.33      P      10.80        A
.      .          .      .           .      .            .
.      .          .      .           .      .            .
.      .          .      .           .      .            .
```

This file has header information and intervening columns between data columns. The "Columns Skipped" setting should be changed to 1 to import this file. Furthermore, the array names are found on row 4. The output files will have the names, "Chip1.ges", "Chip2.ges", "Chip3.ges", etc.

File Specifications

Delimiter	Tab
Columns Skipped	1
Column Names Row	4
Gene Names Column	1
First Data Column	Gene Names Column + 1

Compatible Example 6: Comma-Delimited Text File with Multiple Intervening Columns

```
Gene,Chip1.cel,SD,Call,Chip2.cel,SD,Call,Chip3.cel,SD,Call
H-1,84.76,0.05,A,111.33,0.42,P,30.65,0.57,A
H-2,13.82,2.91,A,23.66,2.13,A,56.80,4.23,A
H-3,48.12,0.81,A,89.43,4.70,A,144.58,1.48,P
H-4,114.59,3.54,P,189.24,0.40,P,80.21,3.53,A
H-5,133.99,4.55,P,107.47,4.60,P,169.93,0.38,P
.
.
.
```

This file is comma-delimited and has two intervening columns between data columns. The "Columns Skipped" setting should be changed to 2 to import this file. Furthermore, the array names are found on row 4. The output files will have the names, "Chip1.ges", "Chip2.ges", "Chip3.ges", etc.

File Specifications

Delimiter	Comma
Columns Skipped	2
Column Names Row	1
Gene Names Column	1
First Data Column	Gene Names Column + 1

Compatible Example 7: Tab-Delimited Text File with Data before the Gene Names Column

```
Date: Oct. 22, 2006

Indiv  Day   Gene  Desc   Chip1.cel   SD     Call  Chip2.cel   SD     Call
1      4     H-1   A      149.99      8.05   P     125.87      8.38   P
1      4     H-2   B      20.64       4.73   A     164.29      2.96   P
1      4     H-3   C      194.71      2.86   P     131.74      2.44   P
1      4     H-4   D      19.81       3.99   A     22.40       3.43   A
1      4     H-5   E      150.31      3.40   P     107.63      3.22   P
.      .     .     .      .           .      .     .           .      .
.      .     .     .      .           .      .     .           .      .
.      .     .     .      .           .      .     .           .      .
```

This file has two intervening columns with the gene names in column 3. Expression data is first encountered in column 5. Furthermore, the array names are found on row 3. The output files will have the names, "Chip1.ges" and "Chip2.ges".

File Specifications

Delimiter	Tab
Columns Skipped	2
Column Names Row	3
Gene Names Column	3
First Data Column	5

Incompatible File Format Examples

Below are examples of incompatible import files with a discussion about what makes the file incompatible with the ***GESS*** Generic Expression Data Import Engine.

Incompatible Example 1: No Gene Names

```
Chip1.cel
149.99
20.64
194.71
19.81
150.31
.
.
.
```

This file does not contain gene names. Gene names are required for the import procedure to function.

Incompatible Example 2: Gene Names Column after the First Data Column

```
Chip1.cel  Gene
149.99     H-1
20.64      H-2
194.71     H-3
19.81      H-4
150.31     H-5
.          .
.          .
.          .
```

The gene names column must be to the left of (less than) the first data column. You should swap the two columns to make this file compatible.

Incompatible Example 3: Multiple-Row Array Names

```
Gene   Chip1.cel
Name   File
H-1    147.33
H-2    94.19
H-3    132.58
H-4    45.23
H-5    67.08
.      .
.      .
.      .
```

If the column names row were set equal to 1, this file would be incompatible. The first data row must immediately follow the column names row. If the column names row were set at 2, this file would result in an output file named, "File.ges".

Incompatible Example 4: Column(s) of Text after the Expression Data

```
Gene   Chip1.cel    Chip2.cel    Chip3.cel     Month
H-1    178.93       7.27         38.75         Jan
H-2    199.85       118.82       58.76         Jan
H-3    157.68       145.29       34.53         Jan
H-4    45.38        181.77       48.37         Jan
H-5    146.90       68.98        184.92        Jan
.      .            .            .             .
.      .            .            .             .
.      .            .            .             .
```

This file has a column of text after the last column of expression data. In order to read all of the data columns, Columns Skipped must be set equal to 0. This will cause the text column to be read in as expression values, causing an error. If the last column were numeric, an output file would be created for that column as though it contained expression values. In this example, the last column should be removed before running the import engine.

Incompatible Example 5: Inconsistent Number of Intervening Columns between Expression Data Columns

```
Gene   Chip1.cel Chip2.cel   Call   Chip3.cel  Call
H-1    135.40    67.90       A      103.23     P
H-2    0.06      9.42        A      106.65     P
H-3    6.15      25.13       A      62.14      A
H-4    78.24     102.89      P      86.59      A
H-5    30.78     176.33      P      10.80      A
.      .         .           .      .          .
.      .         .           .      .          .
.      .         .           .      .          .
```

This file has columns of text between Chip2.cel and Chip3.cel, but not between Chip1.cel and Chip2.cel. The number of columns to be skipped must be consistent for the entire file. The two text columns should be removed or a column should be added between Chip1.cel and Chip2.cel before importing this file.

Incompatible Example 6: Text Appended at the End of the File

```
Gene   Chip1.cel Call   Chip2.cel   Call   Chip3.cel  Call
H-1    135.40    P      67.90       A      103.23     P
H-2    0.06      A      9.42        A      106.65     P
H-3    6.15      A      25.13       A      62.14      A
H-4    78.24     A      102.89      P      86.59      A
H-5    30.78     A      176.33      P      10.80      A
.      .         .      .           .      .          .
.      .         .      .           .      .          .
.      .         .      .           .      .          .
Date: Feb. 20, 2006
CDF File: Test
Technician: Katelyn
```

This file has additional non-data text at the end of the file. The file to import must contain nothing below the last row of data. The appended information should be removed before importing this file.

Entering Text Files

This section describes how file names are entered into the spreadsheet in preparation for importing text files. Two variables (columns) are required for the import process: one designates the input files and the other receives the .ges file names for further analysis. Any name (entered by clicking on the Variable Info tab at the bottom of the spreadsheet) may be used for these variables.

Input File Names Variable

The Input File Names Variable is a column on the spreadsheet containing a list of filenames and paths of the text files that are to be imported. This variable is required to run the Expression Data Import Engine.

To enter a text file name and path into the spreadsheet:

1. Highlight an empty cell.
2. Either type in the path and file name directly or hit **F7** to browse for an appropriate text file.
3. Repeat steps 1 and 2 until all text files have been entered.

Output File Names Variable

When the import engine is run, a new set of files is generated automatically for use in statistical analyses. These output files have the extension ".ges", and are stored in the folder specified under Folder in which Output Files will be Stored. The path and name of each .ges output file is placed on the spreadsheet in the column specified by the Output File Names Variable. This column of pre-processed output files will become your input files for further statistical analyses. This variable is required to obtain output for statistical analysis.

To specify an output folder under Folder in which Output Files will be Stored:

1. Click on the **Browse** button to the right of the window.
2. Select an output folder and click **OK**.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options specify the variables and major file-importing options that will be used in this procedure.

Input File Specifications

These options are used specify the text files that are to be imported.

Input File Names Variable

Select the variable that contains the list of delimited text files to import. These are generic text files containing expression values from a microarray experiment. Common files that can be imported using this procedure are .txt, .dat, .csv, and .dcp (dChip) expression files. Affymetrix .chp files should be imported using the Affymetrix CHP File Import Engine.

All files included in the list should have the same general file format, i.e., the same delimiter, intervening column structure, row containing column names, column containing gene names, and first data column (See Input File Specifications above). You may import files containing different numbers of genes and/or arrays as long as the file structure is the same (See the tutorial below for an example).

Delimiter

Enter the delimiter that separates columns in the expression input files. All files must delimited by this character. The options are tab, comma, blank space, or semicolon. If blank space is chosen, multiple spaces are treated as a single space.

Columns Skipped

Specify the number of columns between expression data columns in the input files. Often, a single input file contains multiple columns of expression data separated by one or more columns containing other information such as standard deviations, presence/absence calls, etc. This other information is ignored by ***GESS*** during one-channel data importing. All files must contain the same number of ignored columns (if any) between expression data columns.

RANGE: Columns Skipped ≥ 0

Column Names Row

Specify the row containing column titles. Column titles (if present) must be on the same row for all input files. The first row of data must immediately follow the column names row. The column titles are used as the basis for the output (.ges) file names. If column titles are absent in any input file, they must be absent in all input files. In the event that column titles are absent, ***GESS*** will create appropriate output (.ges) file names based on the input file names and data column numbers.

RANGE: Column Names Row ≥ 0

Gene Names Column

Specify the column containing gene names or IDs. Further analyses will use these names to identify genes. This column of gene names must be in the same location for all input files and located to the left of the first column of gene expression data.

RANGE: First Data Column > Gene Names Column ≥ 1

First Data Column

Specify the first column containing expression data. The first column of expression values must be the same for all input files and must be located to the right of the column containing gene (probeset) names.

RANGE: First Data Column > Gene Names Column ≥ 1

GES Output File Specifications

These options are used to specify the storage location and naming of the output .ges files.

Folder in which Output Files will be Stored

Enter the path and name of the folder in which the newly created .ges output files will be stored. The path may be typed directly, or the Browse button may be used to locate the desired folder.

In the default path "%mydocs_NCSS%\DATA\GESS", the designator "%mydocs_NCSS%" represents the path to the ***GESS*** personal data folder (commonly *C:\...\[My] Documents\NCSS \NCSS 2007*).

Output File Names Variable

Select the variable in which the path and names of the newly created .ges files will be stored for future statistical analyses. The path and folder for these .ges files is entered under Folder in which Output Files will be Stored.

A new file (to be used for statistical analyses) will be created for each input array when the procedure is run.

CAUTION: If a variable containing data is entered, the data on the spreadsheet will be overwritten and lost.

Append to File Names

A sequence of characters may be entered here to append to the name of the created file.

For example, if a column of expression values has the title "Slide1.cel" and "log" is entered here, the newly created .ges file will be "Slide1 log.ges".

If nothing is entered here, the output file names will be the same column (or input file) names with the extension ".ges".

Overwrite existing output (.ges) files with new output (.ges) files

Check this box to overwrite the existing .ges files when files of the same name are encountered.

If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .ges file.

For example, if "Slide1.ges" has already been created and a new "Slide1.ges" file is to be written, the new file will be "Slide1 (2).ges" if the Overwrite box is not checked.

Transformation Options

These options determine the expression value that will be stored in the output files.

Data Transformation

Specify the transformation to perform on the input expression data before saving output files. This option allows you to replace negative numbers and/or perform logarithmic transformations on the data. If a log transformation is chosen, negative numbers and zeros in the data must be replaced by either missing values or the Replacement Value before the log transformation is computed. The available options are:

- **None**

  No transformation is performed on the data.

- **Set Negative Numbers to Zero**

  Negative numbers are replaced with zeros. No log transformation is performed.

- **Set Negative Numbers to the Replacement Value**

  Negative numbers are replaced with the Replacement Value. No log transformation is performed.

- **Set Negative Numbers to the Missing Values**

  Negative numbers data are replaced with missing values. No log transformation is performed.

- **Log Base X (Set Negative Numbers and Zeros to the Replacement Value)**

  Negative numbers and zeros are replaced with the Replacement Value prior to taking the log of the data. The log options are log base 2, log base e (natural log), and log base 10.

- **Log Base X (Set Negative Numbers and Zeros to the Missing Values)**

  Negative numbers and zeros are replaced with missing values prior to taking the log of the data. The log options are log base 2, log base e (natural log), and log base 10.

Replacement Value

Specify the value that will replace negative values and/or zeros in the dataset. This option is only used when the "Data Transformation" selected requires the use of the Replacement Values. Otherwise, this option is ignored.

RANGE: Replacement Value > 0

Reports Tab

The options on this panel control which reports and plots are generated.

Select Reports

The following reports and report options are available.

File Processing Summary

Check this box to obtain a row-by-row summary of input and output file names and a summary of input and output file specifications.

Data Summary

Check this box to obtain a numeric summary of the data saved in the newly created .ges files.

Decimals

Specify the number of decimals to be used for percentiles in the Data Summary report. The number of decimal places is used for output display only. It does not change the internal precision of the data.

Select Plots

Choose from the following plots.

Comparative Box Plot

Check this box to obtain a comparative box plot of expression values.

Box Plot Tab

The options on this panel control attributes of the box plots.

Horizontal and Vertical Axes

The following options allow you to format the horizontal (X) and vertical (Y) axes.

Label

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate X-axis label.

{Y} is replaced by an appropriate Y-axis label.

Ref. Number Format...

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Minimum

Specify the value to be displayed as the minimum on this axis. If this value is left blank, the minimum will be determined from the data. If this value is greater than the smallest 10th percentile of the data, it will be ignored.

Maximum

Specify the value to be displayed as the maximum on this axis. If this value is left blank, the maximum will be determined from the data. If this value is less than the largest 90th percentile of the data, it will be ignored.

Major Ticks

Specify the number of large tickmarks and optional grid lines along the axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along the axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along the axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the corresponding axis. Thus, checking Show Grid Lines here will actually cause horizontal grid lines to appear.

Box Plot Settings

The following options allow you to control the appearance of the box plot.

Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created in the Box Plots procedure.

% Space

When the Box Width (or Bar Width) option is set to Percent Space in the box plot style file selected, this value specifies the percent of the length of the axis that is empty space instead of bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also, note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**

  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**

  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Interior

The color used to fill the rectangle formed by the vertical and horizontal axes. Click to change.

Background

The color used behind the plot. Click to change.

Box Fill

The color used to fill the boxes. Click to change.

Box Border

The color used to outline the boxes. Click to change.

Line

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Top and Bottom Titles

Enter text here for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate horizontal-axis label.

{Y} is replaced by an appropriate vertical-axis label.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Importing Multiple Expression Data Files (One Column of Data per File)

This section presents an example of how to import six expression files, each having only one column of expression data.

The spreadsheet data used are recorded in the DATAIMPORT1 dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\DataImport* directory. Below is a sample from one of the data files.

```
Gene        Chip1.cel
Gene-1      46.90135409
Gene-2      115.847121
Gene-3      141.8949366
Gene-4      156.7315653
Gene-5      196.8725591
.           .
.           .
.           .
```

To run this example, take the following steps or load the **Example 1** template on the Generic Expression Data Import Engine Template tab.

1 Open the DATAIMPORT1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **DATAIMPORT1.S0**.
- Click **Open**.

2 Open the Generic Expression Data Import Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Generic Expression Data Files**. The Generic Expression Data Import Engine procedure will be displayed.
- On the Generic Expression Data Import Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Generic Expression Data Import Engine window, select the **Variables tab**.
- Under Input file Specifications, set the **Input File Names Variable** to **Infiles**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **Outfiles**.
- Put a check next to **Overwrite existing output (.ges) with new output (.ges) files**.
- Leave all other options under the Variables tab at their default settings.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Input File Summary

```
Input File Summary

        Data
Row     Columns      File Name
1       1            ...\DATA\GESS\DataImport\DataImport_Ex1_1.txt
2       1            ...\DATA\GESS\DataImport\DataImport_Ex1_2.txt
3       1            ...\DATA\GESS\DataImport\DataImport_Ex1_3.txt
4       1            ...\DATA\GESS\DataImport\DataImport_Ex1_4.txt
5       1            ...\DATA\GESS\DataImport\DataImport_Ex1_5.txt
6       1            ...\DATA\GESS\DataImport\DataImport_Ex1_6.txt

Input File Specifications

Parameter                    Value
Input File Names Variable    Infiles
Number of Input Files        6
Column Names Row             1
Gene Names Column            1
First Data Column            2
Columns Skipped              0
Delimiter                    Tab
```

This report displays a list of input files and the number of expression data columns contained in each input file. This report also lists the input file specifications used.

Row

This is the row of the input file on the spreadsheet.

Data Columns

This is the number of expression data columns stored in each input file.

File Name

This contains the path and name of each input file.

Output File Summary

```
Output File Summary

        Number of    Output         Column         Parent (Input)
Row     Genes        File Name      Name           File Name
1       25           Chip1.ges      Chip1.cel      DataImport_Ex1_1.txt
2       25           Chip2.ges      Chip2.cel      DataImport_Ex1_2.txt
3       25           Chip3.ges      Chip3.cel      DataImport_Ex1_3.txt
4       25           Chip4.ges      Chip4.cel      DataImport_Ex1_4.txt
5       25           Chip5.ges      Chip5.cel      DataImport_Ex1_5.txt
6       25           Chip6.ges      Chip6.cel      DataImport_Ex1_6.txt

Output File Specifications

Parameter                     Value
Output File Names Variable    Outfiles
Number of Output Files        6
Output File Folder            ...\DATA\GESS
Data Transformation           None
```

This report displays a list of the output file names and the number of genes contained in each output file. A list of output file specifications is also given.

Row

This is the row of the newly created output file on the spreadsheet. If an input file contains data from more than one array, the output file row may not equal the corresponding input file row.

Number of Genes

This is the number or genes contained in each output file.

Output File Name

These are the names of the newly created output (.ges) files. The folder into which these files were stored is listed as the "Output File Folder".

Column Name

These are the column names encountered in the input files that were used to create the output file names. If a column name is not found, ***GESS*** uses the input file name and column number to create the new output file name.

Parent (Input) File Name

These are the input files from which each output file was created. The path to each file can be found in the Input File Summary.

Numerical Summary of Expression Values

```
Numerical Summary of Expression Values

                          10th        25th        50th        75th        90th
Row    Minimum      Percentile  Percentile  Percentile  Percentile  Percentile     Maximum
1     20.08366        24.25456    73.69063    115.8471    168.8725    190.6173    196.8726
2     18.61828        55.83778    110.1334    154.3349    283.3263     330.777    365.5356
3      14.6164        34.15271    62.92199    108.0773    145.7818    175.9546    196.7924
4     14.31315        29.22659    61.84759    122.0294    197.3421    221.8619     248.026
5      17.4746        21.98588    55.99351    137.3648    197.3245     217.108    240.0826
6     8.134866        24.05448    97.88805    199.4747    358.4406    381.6055    388.5171
```

This report gives numerical summaries of the expression values saved in the output files.

Row

This is the row of the output file on the spreadsheet.

Minimum

This is the minimum expression value stored.

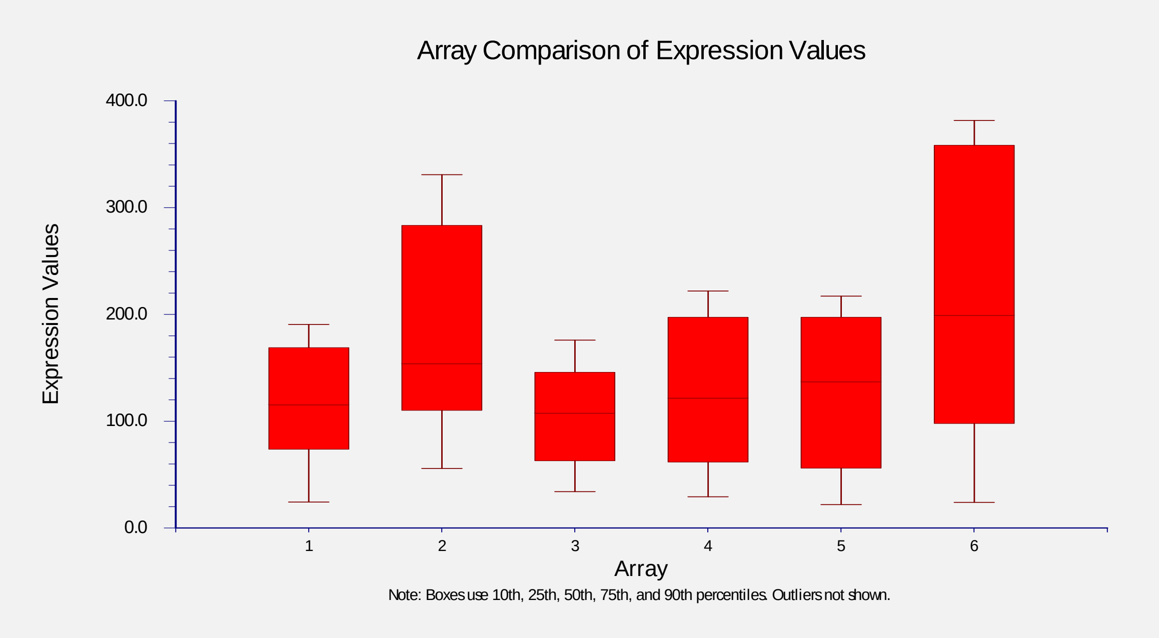
Percentiles

These are the 10th, 25th, 50th, 75th, and 90th expression value percentiles.

Maximum

This is the maximum expression value stored.

Array Comparison of Expression Values



This plot shows the relative distributions of expression values.

Example 2 – Importing Multiple Expression Files (More than One Column of Data per File)

This section presents an example of how to import six expression files, each having only one column of expression data.

The spreadsheet data used are recorded in the DATAIMPORT2 dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\DataImport* directory. Below is a sample from each of the data files. Notice that the second file not only has more columns than the first, but the gene names are entirely different. However, the file format is the same for both input files so they can be imported at the same time.

MAImport_Ex2_1.txt

```
Gene      Chip1.cel  Call   Chip2.cel  Call   Chip3.cel  Call
Gene-1    98.99      A      16.72      A      287.64     P
Gene-2    112.58     P      192.59     P      134.28     P
Gene-3    125.67     P      247.79     P      288.69     P
Gene-4    37.25      A      12.25      A      256.81     P
Gene-5    25.19      A      22.73      A      188.47     P
.         .          .      .          .      .          .
.         .          .      .          .      .          .
.         .          .      .          .      .          .
```

MAImport_Ex2_2.txt

```
Gene      Trt1.cel   Call   Trt2.cel   Call   Trt3.cel   Call   Trt4.cel   Call
Human-1   10.85      A      188.87     P      125.11     P      212.30     P
Human-2   120.83     P      175.52     P      145.72     P      33.58      A
Human-3   104.89     P      140.60     P      101.10     P      142.80     P
Human-4   90.21      A      46.64      A      7.55       A      111.05     P
Human-5   140.38     P      91.99      A      230.95     P      45.33      A
.         .          .      .          .      .          .      .          .
.         .          .      .          .      .          .      .          .
.         .          .      .          .      .          .      .          .
```

To run this example, take the following steps or load the **Example 2** template on the Generic Expression Data Import Engine Template tab.

1 Open the DATAIMPORT2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **DATAIMPORT2.S0**.
- Click **Open**.

2 Open the Generic Expression Data Import Engine window.

- On the menus, select **GESS**, then **Import Microarray Data**, then **Generic Expression Data Files**. The Generic Expression Data Import Engine procedure will be displayed.
- On the Generic Expression Data Import Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Microarray Expression Data Import Engine window, select the **Variables tab**.
- Under Input file Specifications, set the **Input File Names Variable** to **Infiles**.
- Set **Columns Skipped** equal to **1**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **Outfiles**.
- Put a check next to **Overwrite existing output (.ges) with new output (.ges) files**.
- Leave all other options under the Variables tab at their default settings.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Input File Summary

```
Input File Summary

          Data
Row       Columns       File Name
1         3             ...\DATA\GESS\DataImport\DataImport_Ex2_1.txt
2         4             ...\DATA\GESS\DataImport\DataImport_Ex2_2.txt

Input File Specifications

Parameter                       Value
Input File Names Variable       Infiles
Number of Input Files           2
Column Names Row                1
Gene Names Column               1
First Data Column               2
Columns Skipped                 1
Delimiter                       Tab
```

The report indicates that the input file on row 2 has contains more data columns than the file on row 1.

Output File Summary

```
Output File Summary

          Number of     Output          Column          Parent (Input)
Row       Genes         File Name       Name            File Name
1         25            Chip1.ges       Chip1.cel       DataImport_Ex2_1.txt
2         25            Chip2.ges       Chip2.cel       DataImport_Ex2_1.txt
3         25            Chip3.ges       Chip3.cel       DataImport_Ex2_1.txt
4         30            Trt1.ges        Trt1.cel        DataImport_Ex2_2.txt
5         30            Trt2.ges        Trt2.cel        DataImport_Ex2_2.txt
6         30            Trt3.ges        Trt3.cel        DataImport_Ex2_2.txt
7         30            Trt4.ges        Trt4.cel        DataImport_Ex2_2.txt

Output File Specifications

Parameter                       Value
Output File Names Variable      Outfiles
Number of Output Files          7
Output File Folder              ...\DATA\GESS
Data Transformation             None
```

The report indicates that the number of genes contained in the first three output files is different from the number of genes contained in the latter four. These input files likely come from different array types. This report allows you to make sure that all the arrays you will compare in later analyses are of the same type.

Example 3 – Performing a Data Transformation During Import

This section presents an example of how to import six expression files and perform a logarithmic transformation on the data before saving the output file.

The spreadsheet data used are recorded in the DATAIMPORT2 dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\DataImport* directory.

To run this example, take the following steps or load the **Example 3** template on the Generic Expression Data Import Engine Template tab.

1 **Open the DATAIMPORT2 dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **DATAIMPORT2.S0**.
- Click **Open**.

2 **Open the Generic Expression Data Import Engine window.**

- On the menus, select **GESS**, then **Import Microarray Data**, then **Generic Expression Data Files**. The Generic Expression Data Import Engine procedure will be displayed.
- On the Generic Expression Data Import Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Microarray Expression Data Import Engine window, select the **Variables tab**.
- Under Input file Specifications, set the **Input File Names Variable** to **Infiles**.
- Set **Columns Skipped** equal to **1**.
- Set the **Folder in which Output Files will be Stored** to **%mydocs_NCSS%\DATA \GESS**.
- Set the **Output File Names Variable** to **Outfiles**.
- Under **Append to File Names**, enter **log2**.
- Put a check next to **Overwrite existing output (.ges) with new output (.ges) files**.
- Under **Data Transformation**, select **Log base 2 (Set Negative Numbers and Zeros to the Replacement Value)**.
- Leave all other options under the Variables tab at their default settings.

4 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Input File Summary

The input file summary is the same as that in Example 2.

Output File Summary

The output file summary is the same as that in Example 2.

Transformation Summary

Transformation Summary
Data Transformation: Log base 2 (Set Negative Numbers and Zeros to the Replacement Value = 0.001)

Row	Number of Negatives Replaced	Number of Zeros Replaced	Output File Name
1	0	0	Chip1 log2.ges
2	0	0	Chip2 log2.ges
3	0	0	Chip3 log2.ges
4	0	1	Trt1 log2.ges
5	0	1	Trt2 log2.ges
6	0	0	Trt3 log2.ges
7	1	1	Trt4 log2.ges

This report is only obtained if a data transformation of some type is performed. The report indicates that one negative number (from Trt4) and three zeros (from Trt1, Trt2, and Trt4) were replaced by 0.001, the replacement value. Furthermore, the log base 2 transformation was used after the negative values and zeros were replaced.

Row

This is the row of the output file on the spreadsheet.

Number of Negatives Replaced

This is the total number of negative values that were encountered and replaced for each array or column of expression data.

Number of Zeros Replaced

This is the total number of zeros that were encountered and replaced for each array or column of expression data.

Output File Name

These are the names of the newly created output (.ges) files. The folder into which these files were stored is listed as the "Output File Folder" in the Output File Summary.

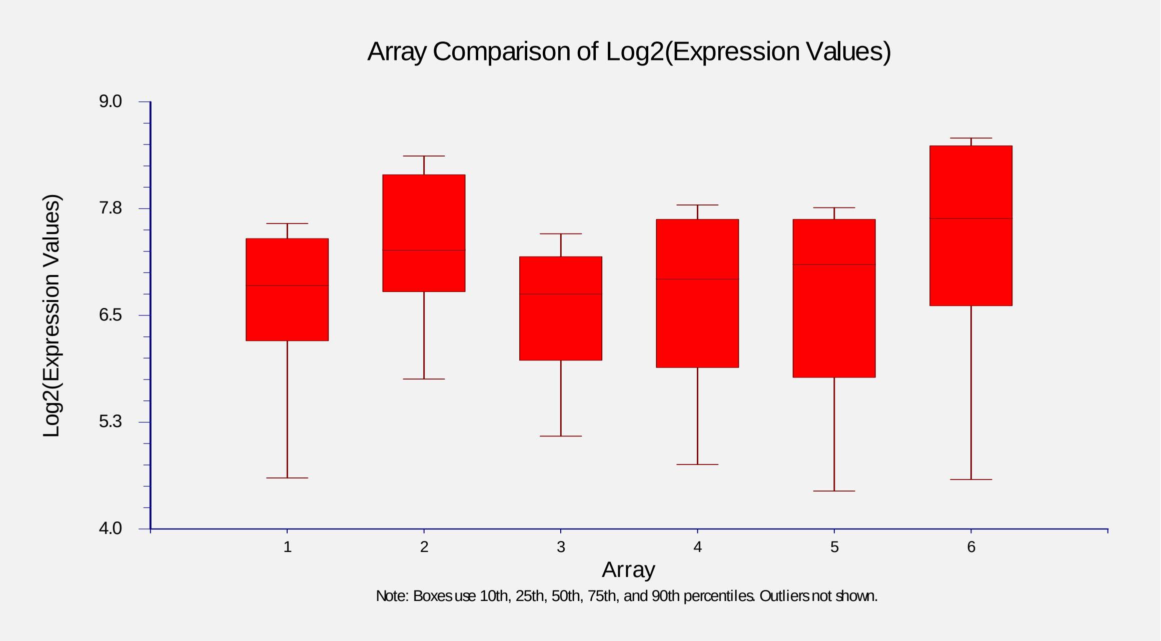
Numerical Summary of Expression Values

Numerical Summary of Log2(Expression Values)

Row	Minimum	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile	Maximum
1	4.654779	5.179523	5.737394	6.486071	7.018985	7.45496	7.545582
2	1.632268	2.344041	4.165145	6.857608	7.46928	7.942536	7.98624
3	1.035624	5.521379	6.79293	7.558191	8.0454	8.177639	8.215678
4	-9.965784	3.414246	4.74773	6.546448	7.117599	7.455993	7.609326
5	-9.965784	4.606921	5.68853	7.263409	7.668276	7.926028	7.966996
6	2.916477	5.086963	6.144235	7.279669	7.907256	8.073811	8.157549
7	-9.965784	3.070641	6.25482	7.2756	7.814419	8.145815	8.170025

These are the summary statistics of the expression values that were stored for each column of expression data. These values are on the log base 2 scale.

Array Comparison of Expression Values



This plot shows the relative distributions of expression values on the log base 2 scale.

Chapter 140

Save Data to Spreadsheet

Introduction

This chapter describes how to save pre-processed expression values to the spreadsheet using the ***GESS*** Save Data to Spreadsheet procedure. Gene lists may be entered directly, from the spreadsheet, or from a separate text file.

Before running this procedure, output (.ges) files containing a single expression value for each gene on each array must be obtained using the appropriate pre-processing procedure in ***GESS***.

Overview

The ***GESS*** Save Data to Spreadsheet procedure allows the user to obtain pre-processed expression values from .ges files for any of the genes of those files. Pre-processed expression values may be stored to the spreadsheet in the hypothesis testing procedures (e.g., ***GESS*** T-Test – Two Groups or ***GESS*** Multiple Regression), but only for those genes for which the significance level falls below a specified cutoff. This procedure gives the user the freedom of obtaining values for genes of specific interest to the researcher, which may or may not be significant in a multiple testing adjusted hypothesis test.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options specify the variables that will be used in this procedure.

Expression Value Storage Variables

This section allows you to specify the variables to be used in this procedure.

GES Files Variable

Enter here the variable containing the list of .ges files from which the pre-processed expression values will be obtained. The .ges files are created using one of the GESS pre-processing procedures.

First Storage Variable

The expression values for the first gene (after alpha-numeric ordering) will be stored in this variable. The values for each additional gene are stored in the variables immediately to the right of this variable.

WARNING: Use caution when selecting this variable, since existing data will be replaced when the storage variables are created.

Specify Genes to Store

Use this section to specify the genes to output to the spreadsheet.

Genes for Expression Storage on the Spreadsheet

Enter the list of genes for which the pre-processed gene expression data of the .ges files will be stored onto the spreadsheet. The genes may be entered directly, or the * character may be used to specify all genes with a particular beginning. A list of genes may be enter as a file, using the notation file(...\...\filename.txt). The file must contain a list of gene names or IDs, each on a separate line. A list of genes may also be entered as a column of the spreadsheet, using the notation var(variable name).

EXAMPLES:

1. Blank

2. spike1
   spike3
   spike5
   spike7

3. spike* (all names beginning with spike)

4. AA44719

5. NM_00582
   NM_04762
   NM_27564

6. cntrl* (all names beginning with cntrl)

7. file(C:\Microarray\genelist.txt) (all names in the genelist.txt file)

8. var(OutputGenes) (all names in the spreadsheet variable with the variable name OutputGenes)

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Saving Data to the Spreadsheet

Suppose the effect of a hormone on gene expression of mice is to be analyzed for 345 genes. A control and three levels of hormone (4%, 8%, and 12%) are to be administered. Twenty-four mice are randomly assigned to the four hormone levels so that there are 6 mice in each treatment. A single cDNA sample is obtained from each experimental unit (mouse) following treatment. Each sample is exposed to a single microarray, resulting in a single expression value for each gene for each unit of each treatment group.

The goal is to determine for each gene whether there is evidence that the expression is different across hormone levels.

Regardless of the significance level, three genes of particular interest are to be examined in detail by outputting the pre-processed data for these genes to the spreadsheet. The genes are

36678_at
39425_at
35201_at

In the pre-processing procedure, 24 files are created. The format of the spreadsheet is shown below. The spreadsheet data used are recorded in the SAVEDATA dataset. The designator "%p%" represents the path to the folder into which ***NCSS*** and ***GESS*** were installed (commonly *C:\Program Files\NCSS\NCSS 2007*).

SAVEDATA dataset

Treatment	OutputFile	Genes
Control	%p%\DATA\GESS\SaveData\SaveData_1.ges	36678_at
Control	%p%\DATA\GESS\SaveData\SaveData_2.ges	39425_at
Control	%p%\DATA\GESS\SaveData\SaveData_3.ges	35201_at
Control	%p%\DATA\GESS\SaveData\SaveData_4.ges	
Control	%p%\DATA\GESS\SaveData\SaveData_5.ges	
Control	%p%\DATA\GESS\SaveData\SaveData_6.ges	
Trt1	%p%\DATA\GESS\SaveData\SaveData_7.ges	
Trt1	%p%\DATA\GESS\SaveData\SaveData_8.ges	
Trt1	%p%\DATA\GESS\SaveData\SaveData_9.ges	
Trt1	%p%\DATA\GESS\SaveData\SaveData_10.ges	
Trt1	%p%\DATA\GESS\SaveData\SaveData_11.ges	
Trt1	%p%\DATA\GESS\SaveData\SaveData_12.ges	
Trt2	%p%\DATA\GESS\SaveData\SaveData_13.ges	
Trt2	%p%\DATA\GESS\SaveData\SaveData_14.ges	
Trt2	%p%\DATA\GESS\SaveData\SaveData_15.ges	
Trt2	%p%\DATA\GESS\SaveData\SaveData_16.ges	
Trt2	%p%\DATA\GESS\SaveData\SaveData_17.ges	
Trt2	%p%\DATA\GESS\SaveData\SaveData_18.ges	
Trt3	%p%\DATA\GESS\SaveData\SaveData_19.ges	
Trt3	%p%\DATA\GESS\SaveData\SaveData_20.ges	
Trt3	%p%\DATA\GESS\SaveData\SaveData_21.ges	
Trt3	%p%\DATA\GESS\SaveData\SaveData_22.ges	
Trt3	%p%\DATA\GESS\SaveData\SaveData_23.ges	
Trt3	%p%\DATA\GESS\SaveData\SaveData_24.ges	

To run this example and output the pre-processed data for the three genes, take the following steps or load the **Example 1** template on the ***GESS*** Save Data to Spreadsheet Template tab.

1 Open the SAVEDATA dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Open the **GESS** folder.
- Click on the file **SAVEDATA.S0**.
- Click **Open**.

2 Open the GESS Save Data to Spreadsheet window.

- On the menus, select **GESS**, then **Data Utilities**, then **Save Data to Spreadsheet**. The ***GESS*** Save Data to Spreadsheet procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables and storage details.

- On the Save Data to Spreadsheet window, select the **Variables tab**.
- Set the **GES Files Variable** to **OutputFile**.

- Set the **First Storage Variable** to **C5**.
- Use one of the following under **Genes for Expression Storage on the Spreadsheet**:
  1. Type in the three gene names directly:

     36678_at

     39425_at

     35201_at
  2. Enter the variable containing the gene names using the var() notation:

     var(Genes)
  3. Enter the file containing the gene names using the file() notation:

     file(%p%\data\gess\SaveData\ThreeGenes.txt)

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Spreadsheet Storage Data

The spreadsheet should now appear as follows:

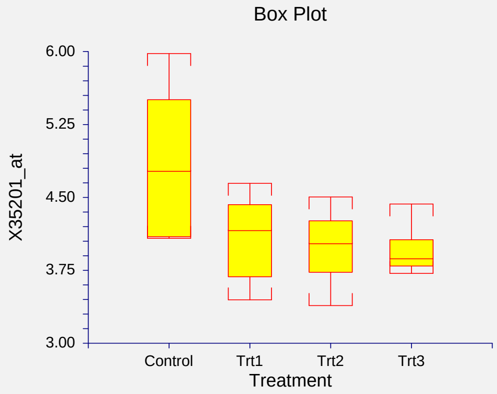
Treat ment	OutputFile	Genes	C4	X35021_at	X36678_at	X39425_at
Control	...Data_1.ges	36678_at		5.345318946	4.629663607	3.776854412
Control	...Data_2.ges	39425_at		4.673518828	3.584157852	4.26761696
Control	…ata_3.ges	35201_at		4.07930048	3.748580026	4.06967838
Control	...Data_4.ges			5.97979466	3.838779192	3.572910858
Control	...Data_5.ges			4.879765374	4.509594926	4.201394181
Control	...Data_6.ges			4.100708841	4.031677811	4.050460081
Trt1	...Data_7.ges			4.352520909	3.706517136	3.751461025
Trt1	…Data_8.ges			3.445652577	4.366076354	5.090037108
Trt1	…Data_9.ges			4.260463808	4.49971818	4.982597403
Trt1	…Data_10.ges			4.072603013	4.019008036	4.124968326
Trt1	…Data_11.ges			4.644646909	4.02867229	3.606809011
Trt1	…Data_12.ges			3.762897025	4.731043161	4.975243228
Trt2	…Data_13.ges			3.908622212	3.93514597	4.177175603
Trt2	…Data_14.ges			3.387230244	3.956506817	3.944409002
Trt2	…Data_15.ges			4.152148588	3.676527218	3.896450862
Trt2	…Data_16.ges			4.177154461	4.853560704	4.511311866
Trt2	…Data_17.ges			4.504194041	4.955175022	3.709320123
Trt2	…Data_18.ges			3.844255517	4.165499617	4.567651517
Trt3	…Data_19.ges			3.717092646	4.781196499	3.702486801
Trt3	…Data_20.ges			3.819677523	5.961919638	4.814098706
Trt3	…Data_21.ges			3.878897141	3.879797164	3.769329829
Trt3	…Data_22.ges			3.941843587	4.165123199	4.308604726
Trt3	…Data_23.ges			3.872591221	3.577611746	3.875418154
Trt3	…Data_24.ges			4.432201814	4.174162566	3.475477871

The pre-processed data in the final three columns can then be used for individual analyses. Hypothesis tests done on the individual genes will not, however, be adjusted for multiplicity of genes.

ANOVA Output

For example, one-way analysis of variance (using the ***NCSS*** One-Way Analysis of Variance procedure) could be run on X35201_at to obtain:

```
Tests of Assumptions Section
                                            Test          Prob          Decision
Assumption                                  Value         Level         (0.05)
Skewness Normality of Residuals             0.7417        0.458244      Accept
Kurtosis Normality of Residuals             0.8286        0.407345      Accept
Omnibus Normality of Residuals              1.2367        0.538828      Accept
Modified-Levene Equal-Variance Test         2.1967        0.120062      Accept

Box Plot Section
```



```
Expected Mean Squares Section
Source                                Term         Denominator     Expected
Term                     DF           Fixed?       Term            Mean Square
A: Treatment             3            Yes          S(A)            S+sA
S(A)                     20           No                           S(A)
Note: Expected Mean Squares are for the balanced cell-frequency case.

Analysis of Variance Table
Source                                Sum of       Mean                       Prob          Power
Term                   DF             Squares      Square         F-Ratio     Level         (Alpha=0.05)
A: Treatment           3              3.191029     1.063676          4.56     0.013702*     0.812576
S(A)                   20             4.668373     0.2334186
Total (Adjusted)       23             7.859402
Total                  24
* Term significant at alpha = 0.05

Kruskal-Wallis One-Way ANOVA on Ranks
Hypotheses
Ho: All medians are equal.
Ha: At least two medians are different.

Test Results
                                                Chi-Square      Prob
Method                      DF                  (H)             Level          Decision(0.05)
Not Corrected for Ties      3                   7.74            0.051702       Accept Ho
Corrected for Ties          3                   7.74            0.051702       Accept Ho

Number Sets of Ties         0
Multiplicity Factor         0
```

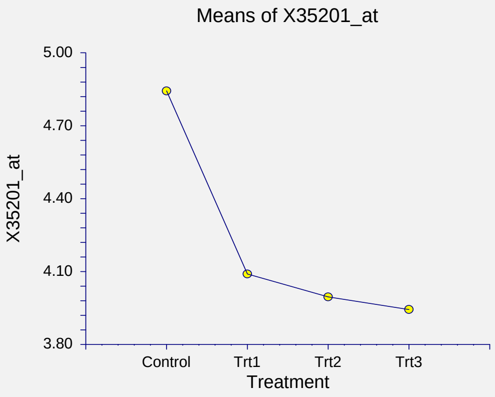
Group Detail

Group	Count	Sum of Ranks	Mean Rank	Z-Value	Median
Control	6	115.00	19.17	2.6667	4.776642
Trt1	6	70.00	11.67	-0.3333	4.166533
Trt2	6	64.00	10.67	-0.7333	4.030385
Trt3	6	51.00	8.50	-1.6000	3.875744

Means and Effects Section

Term	Count	Mean	Standard Error	Effect
All	24	4.218046		4.218046
A: Treatment				
Control	6	4.843068	0.1972387	0.625022
Trt1	6	4.089797	0.1972387	-0.1282485
Trt2	6	3.995601	0.1972387	-0.222445
Trt3	6	3.943717	0.1972387	-0.2743285

Plots of Means Section



Tukey-Kramer Multiple-Comparison Test

Response: X35201_at
Term A: Treatment

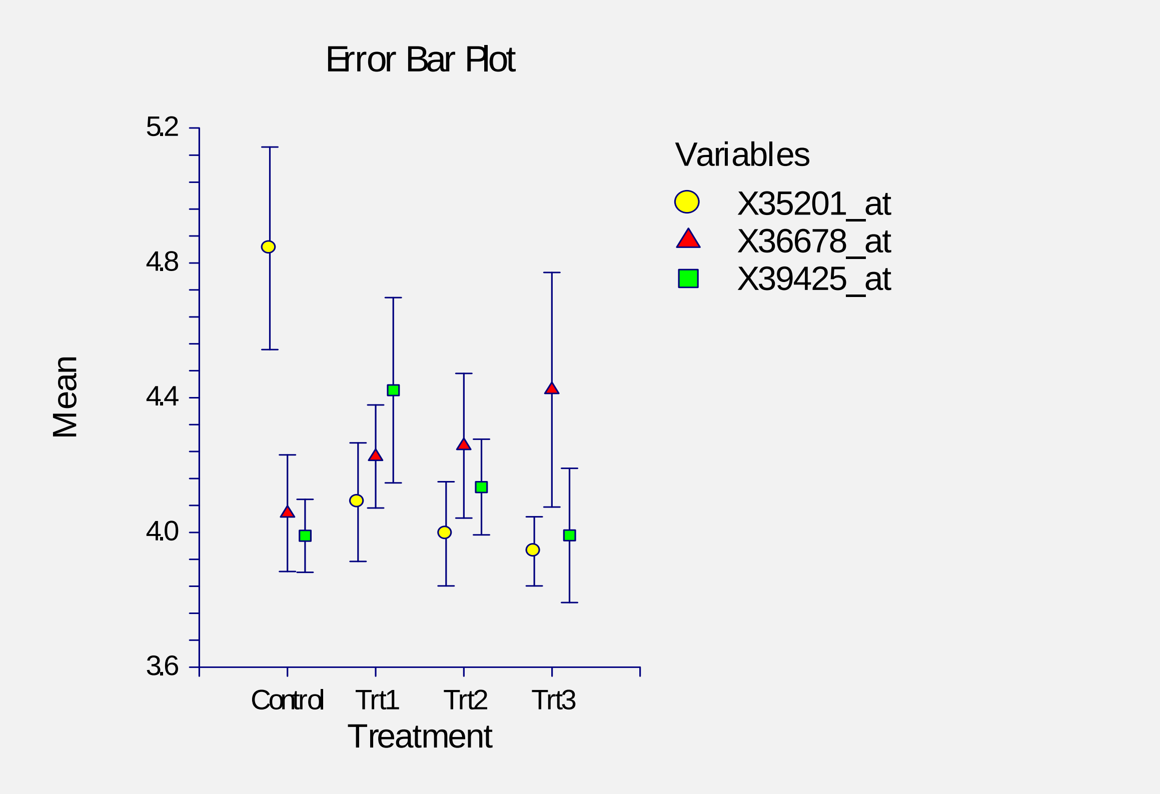
Alpha=0.050 Error Term=S(A) DF=20 MSE=0.2334186 Critical Value=3.9583

Group	Count	Mean	Different From Groups
Trt3	6	3.943717	Control
Trt2	6	3.995601	Control
Trt1	6	4.089797	
Control	6	4.843068	Trt3, Trt2

Notes:
This report provides multiple comparison tests for all pairwise differences between the means.

Error-Bar Chart

Error-bar charts may be obtained for the three genes using the ***NCSS*** Error-Bar Charts procedure:



Several other analysis and graphics procedures could be used to analyze and describe the data for each of the genes.

Chapter 145

GES File Description

Introduction

This chapter explains how to obtain file information from binary .ges files created using ***GESS***. Several options are available during pre-processing and importing of expression data. These options along with creation and parent file information are stored in each .ges file. This procedure allows you to retrieve and view important information stored in .ges files.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This tab specifies the variable and options that will be used in this procedure.

GES File Specifications

These options are used specify the .ges files that are to be described.

GES File Names Variable

Select the variable that contains the list of .ges files to be described. The names and paths of the files should appear in a column below this variable name on the spreadsheet. To enter the .ges file names and paths into the spreadsheet:

1. Highlight an empty cell.
2. Either type in the path and file name directly or hit F7 to browse for an appropriate .ges file.
3. Repeat steps 1 and 2 until all .ges files have been entered.

Reports Tab

The options on this panel control which reports and plots are generated.

File Reports

Choose from the following file summary reports.

File Summary

Check this box to obtain a description of each .ges file.

Parent (Input) File Summary

Check this box to obtain a description of the parent file corresponding to each .ges file.

Processing Options Summary

Check this box to obtain a summary of the processing options used during the creation of each .ges file.

Files Processed Summary (Affymetrix CEL Only)

Check this box to print out the list of .cel files that were processed together using the ***GESS*** Affymetrix CEL File Pre-Processing Engine. This option is only used for .ges files that were created from Affymetrix .cel files using RMA.

Parent (Input) File Header

Check this box to view the header from the parent file corresponding to each .ges file.

Numeric Reports

Choose from the following numeric summary reports and options.

Data Summary

Check this box to obtain a numeric summary of expression values saved in each .ges file.

Decimals

Specify the number of decimals to be used for percentiles in the Data Summary report. The number of decimal places is used for output display only. It does not change the internal precision of the data.

Plots

Choose from the following plots.

Box Plot

Check this box to obtain a comparative box plot of expression values.

Box Plot Tab

The options on this panel control attributes of the box plots.

Horizontal and Vertical Axes

The following options allow you to format the horizontal (X) and vertical (Y) axes.

Label

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate X-axis label.

{Y} is replaced by an appropriate Y-axis label.

Ref. Number Format...

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Minimum

Specify the value to be displayed as the minimum on this axis. If this value is left blank, the minimum will be determined from the data. If this value is greater than the smallest 10th percentile of the data, it will be ignored.

Maximum

Specify the value to be displayed as the maximum on this axis. If this value is left blank, the maximum will be determined from the data. If this value is less than the largest 90th percentile of the data, it will be ignored.

Major Ticks

Specify the number of large tickmarks and optional grid lines along the axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along the axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along the axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the corresponding axis. Thus, checking Show Grid Lines here will actually cause horizontal grid lines to appear.

Box Plot Settings

The following options allow you to control the appearance of the box plot.

Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created in the Box Plots procedure.

% Space

When the Box Width (or Bar Width) option is set to Percent Space in the box plot style file selected, this value specifies the percent of the length of the axis that is empty space instead of bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also, note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**

  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**

  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Interior

The color used to fill the rectangle formed by the vertical and horizontal axes. Click to change.

Background

The color used behind the plot. Click to change.

Box Fill

The color used to fill the boxes. Click to change.

Box Border

The color used to outline the boxes. Click to change.

Line

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Top and Bottom Titles

Enter text here for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate horizontal-axis label.

{Y} is replaced by an appropriate vertical-axis label.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Obtaining a GES File Description

This section presents an example of how to get description information from three .ges files. The spreadsheet data used are recorded in the GESFILES dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\Utilities* directory.

To run this example, take the following steps or load the **Example 1** template on the ***GESS*** GES File Description Template tab.

1 Open the GESFILES dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **GESFILES.S0**.
- Click **Open**.

2 Open the GES File Description window.

- From the menus, select **GESS**, then **Data Utilities**, then **GES File Description**. The GES File Description procedure will be displayed.
- On the GES File Description window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the GES File Description window, select the **Variables tab**.
- Under GES File Specifications, set the **GES File Names Variable** to **GES_Files**.

4 Specify the reports.

- On the GES File Description window, select the **Reports tab**.
- Under **File Reports**, put a check next to **Parent (Input) File Header** and **Files Processed Summary**, in addition to the boxes already checked.
- Leave all other options under the Reports tab and other tabs at their default settings.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

A separate set of description reports is generated for each .ges file. The report from the first example file is given below.

File Summary

```
GES File              ...\DATA\GESS\Utilities\SampleGES_1.ges   (Row = 1)

File Summary

Parameter             Value
File Date                                         4/12/2006 3:32:12 PM
Number of Genes       345
Fingerprint           52812
Database              ...\DATA\GESS\GEStest.S0
GES Files Variable    OutputFile
```

This report displays the .ges file summary. This report may change depending on the type of parent file. The most common report items are described below.

File Date

This is the date and time that the .ges file was created.

Number of Genes

This is the number of genes for which expression values are stored in the .ges file.

Fingerprint

This is a number generated by ***GESS*** that is unique for this array type. The fingerprint is used to ensure that comparisons are made between arrays of the same type.

Database

This is the name of the database containing the list of .ges files of which the current .ges file is a member. The list is generated when pre-processing is performed or when a file is imported.

GES Files Variable

This is the name of the variable within the database that contains the current .ges file.

Parent (Input) File Summary

```
Parent (Input) File Summary

Parameter           Value
Parent File Type    Affymetrix CEL File
CEL File            D:\0A70\DATA\GESS\SampleGES_1.cel
CDF File            D:\0A70\DATA\GESS\AF\Test3.cdf
CDH File            D:\0A70\DATA\GESS\CDH\Test3.cdh
Database            D:\0A70\DATA\GESS\GEStest.S0
Input Variable      InputFile
```

This report displays a summary of the parent file. The most common report items are described below.

Parent File Type

This is the type of file that was used in the creation of the current .ges file.

Database

This is the name of the database with the list of .ges files of which the current .ges file is a member. The list is generated when pre-processing is performed or when a file is imported.

Input Variable

This is the name of the variable within the database that contains the parent file that was used to create the current .ges file.

Processing Options Summary

```
RMA Pre-Processing Options Summary

Parameter               Value
Background Correction   RMA (Model-Based)
Normalization           Quantile Normalization
Summarization           Median Polish
Output Scale            Log base 2
```

This report presents the options that were used in pre-processing or importing. The items on the report depend entirely on the type of file imported.

Files Processed Summary (Affymetrix CEL Only)

```
Processed Files Summary

Parameter                  Value
Database                   D:\0A70\DATA\GESS\GEStest.S0
Input Variable             InputFile
Number of Files Processed  6

Files Used in RMA Pre-Processing
D:\0A70\DATA\GESS\SampleGES_1.cel
D:\0A70\DATA\GESS\SampleGES_2.cel
D:\0A70\DATA\GESS\SampleGES_3.cel
D:\0A70\DATA\GESS\SampleGES_4.cel
D:\0A70\DATA\GESS\SampleGES_5.cel
D:\0A70\DATA\GESS\SampleGES_6.cel
```

This report summarizes the files that were processed together using RMA to create the expression values contained in the .ges file.

Database

This is the name of the database with the list of .ges files of which the current .ges file is a member. The list is generated when pre-processing is performed or when a generic file is imported.

Input Variable

This is the name of the variable within the database that contains the .cel files that were processed together to create the current .ges file.

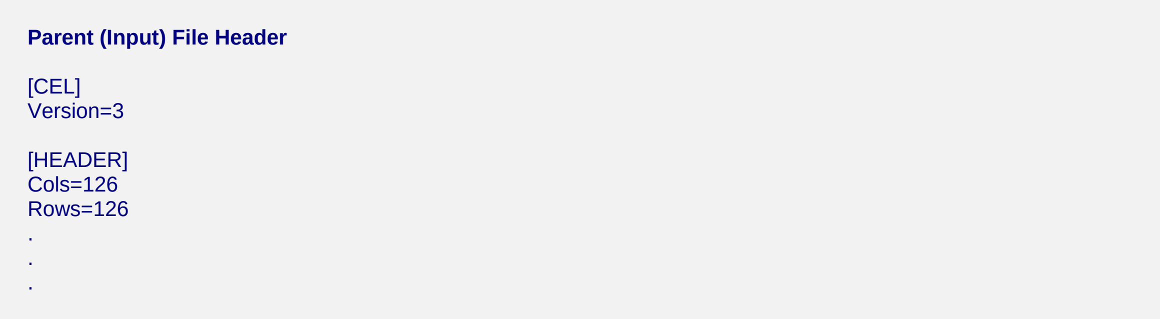
Number of Files Processed

This is the number of .cel files that were processed together to create the current .ges file.

Files Used in RMA Pre-Processing

These are the names of the files that were pre-processed together using RMA to create the current .ges file.

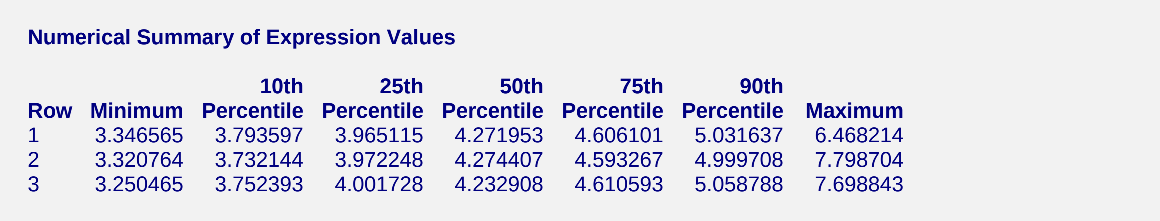
Parent (Input) File Header



```
Parent (Input) File Header

[CEL]
Version=3

[HEADER]
Cols=126
Rows=126
.
.
.
```

This report displays the header pulled from the parent file at the time that the current .ges file was created. The content depends entirely on the type of file.

Numerical Summary of Expression Values



```
Numerical Summary of Expression Values

                   10th        25th        50th        75th        90th
Row  Minimum  Percentile  Percentile  Percentile  Percentile  Percentile   Maximum
1    3.346565   3.793597    3.965115    4.271953    4.606101    5.031637  6.468214
2    3.320764   3.732144    3.972248    4.274407    4.593267    4.999708  7.798704
3    3.250465   3.752393    4.001728    4.232908    4.610593    5.058788  7.698843
```

This report gives summary statistics of the expression values saved in the .ges files.

Row

This is the row of the .ges file on the spreadsheet.

Minimum

This is the minimum expression value.

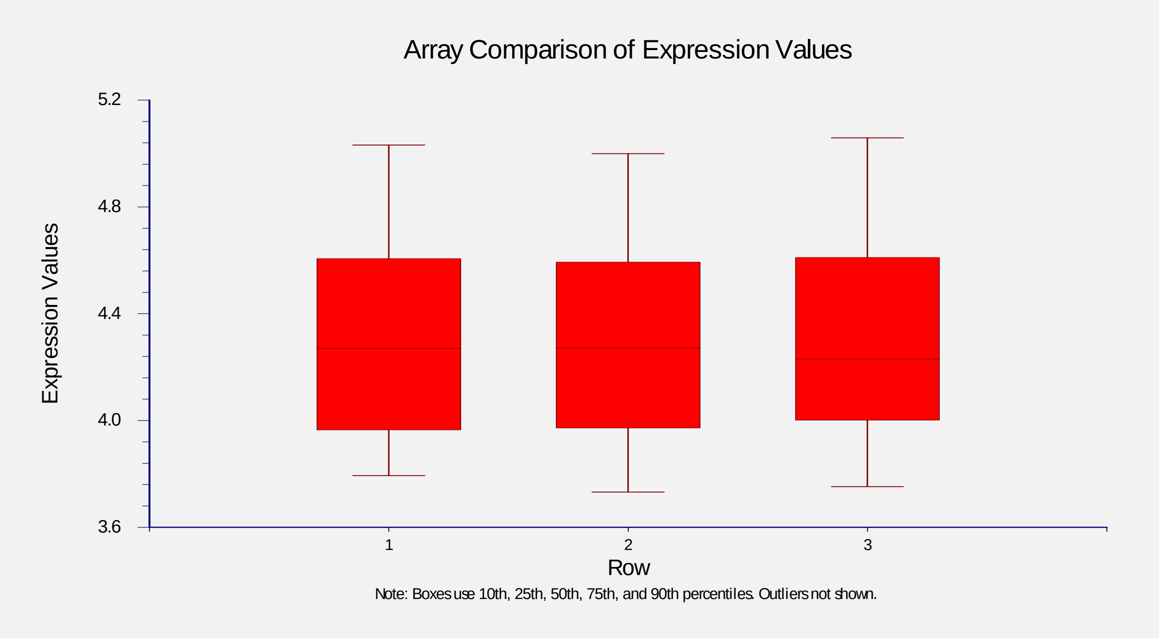
Percentiles

These are the 10th, 25th, 50th, 75th, and 90th expression value percentiles.

Maximum

This is the maximum expression value.

Array Comparison Box Plot



This plot shows the relative distributions of expression values.

145-10 GES File Description

Chapter 150

Data Export Engine

Introduction

This chapter describes the process of exporting data from .ges files created using ***GESS*** to generic text (.txt) files using the ***GESS*** Data Export Engine. The .ges files used by ***GESS*** are created as binary files and can not be read by other applications without first exporting the data to text files. Multiple files may be exported on a single run of the ***GESS*** Data Export Engine. The engine gives you the ability to export each .ges file to a separate text file or to export all .ges files to a single delimited text file containing all of the expression data. If you choose to export to multiple text files, you have the option of exporting important file processing information to the text files along with the expression values. The exported text files can be imported into other expression analysis software as desired.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options specify the variables and major file-exporting options that will be used in this procedure.

GES File Specifications

These options are used specify the .ges files that are to be exported.

GES File Names Variable

Select the variable that contains the list of .ges files to export. The names and paths of the files should appear in a column below this variable name on the spreadsheet. To enter the .ges file names and paths into the spreadsheet:

1. Highlight an empty cell.
2. Either type in the path and file name directly or hit F7 to browse for an appropriate .ges file.
3. Repeat steps 1 and 2 until all .ges files have been entered.

TXT Output File Specifications

These options are used to specify the type, location and naming of the output .txt files.

Export Type

This option allows you to select the type of export operation to perform. This option controls the number of output files that will be created. The choices are

- **Multiple Files**

  Selecting this option forces the creation of a separate output text file for each .ges file. The files are stored in the location specified under Folder in which Multiple Text Files will be Stored. The output file names will be based on the .ges file names. For example, if a .ges file has the name "Slide1.ges", the exported text file will be "Slide1.txt".

- **Single File**

  Selecting this option forces the creation of only one output text file with the data from all .ges files. In order to create a single output file, all .ges files must be compatible, i.e., they must contain expression data for the same gene list. The path and name of the single output text file is specified under Single Text File Name. The Folder in which Multiple Text Files will be Stored, Append to File Names, Overwrite, and Save options are ignored if this option is checked.

Text File Delimiter

Enter the delimiter that will separate columns in the output expression file(s). The options are Tab, Comma, Space, and Semicolon.

Multiple Files Storage Options

The following options are only used if Export Type is set to "Multiple Files".

Folder in which Multiple Text Files will be Stored

Enter the path and name of the folder in which the newly created .txt files will be stored if Export Type is set to "Multiple Files". The path may be typed directly or the browse button to the right may be used to locate the desired folder.

In the default path "%mydocs_NCSS%\DATA\GESS", the designator "%mydocs_NCSS%" represents the path to the ***GESS*** personal data folder (commonly *C:\...\[My] Documents\NCSS \NCSS 2007*).

Overwrite existing output (.txt) files with new output (.txt) files

Check this box to overwrite the existing .txt files when files of the same name are encountered. If the box is not checked, and the same name is encountered when writing files, a number will be appended to the name of the newly created .txt file. For example, if Slide1.txt has already been created and a new Slide1.txt file is to be written, the new file will be Slide1 (2).txt if the Overwrite box is not checked.

Append to File Names

A sequence of characters may be entered here to append to the name of the created file. For example, if the .ges file has the name "Slide1_10hours.ges" and "log" is entered here, the newly created .txt file will be "Slide1_10hours log.txt". If nothing is entered here, the new file name will be the same as the name of the .ges file, but ".ges" will be replaced with ".txt".

Save GES File Info

Check this box to save general information about each .ges file.

Save Parent (Input) File Info

Check this box to save general information about the parent file corresponding to each .ges file.

Save Processing Options

Check this box to save the processing options that were used during the creation of each .ges file.

Save Files Processed List (Affymetrix CEL Only)

Check this box to save the list of .cel files that were processed together using the ***GESS*** Affymetrix CEL File Pre-Processing Engine. This option is only used for .ges files that were created from Affymetrix .cel files using RMA.

Save Parent (Input) File Header

Check this box to save the file header from the parent file corresponding to each .ges file.

Single File Storage Options

The following option is only used if Export Type is set to "Single File".

Single Text File Name

This option allows you to specify the name of the single output text file that is created if Export Type is set to "Single File". The path and file name may be typed directly or the browse button to the right may be used to select the desired file name.

Reports Tab

The options on this panel control which reports and plots are generated.

Select Reports

The following reports and report options are available.

File Processing Summary

Check this box to obtain a row-by-row summary of input and output file names and a summary of input and output file specifications.

Data Summary

Check this box to obtain a numeric summary of the data saved in the newly created .txt files.

Decimals

Specify the number of decimals to be used for percentiles in the Data Summary report. The number of decimal places is used for output display only. It does not change the internal precision of the data.

Select Plots

Choose from the following plots.

Comparative Box Plot

Check this box to obtain a comparative box plot of expression values.

Box Plot Tab

The options on this panel control attributes of the box plots.

Horizontal and Vertical Axes

The following options allow you to format the horizontal (X) and vertical (Y) axes.

Label

Enter text here for the designated label.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate X-axis label.

{Y} is replaced by an appropriate Y-axis label.

Ref. Number Format...

This option specifies the characteristics of the reference numbers. It displays a window that edits the font size and color of the reference numbers that appear next to the text along the axis of the plot. It also allows you to set the number of digits in the reference numbers as well as their vertical/horizontal orientation.

Note that in some cases, the format specified here is overridden by the variable's format as specified on the database in the Variable Info Sheet.

Minimum

Specify the value to be displayed as the minimum on this axis. If this value is left blank, the minimum will be determined from the data. If this value is greater than the smallest 10th percentile of the data, it will be ignored.

Maximum

Specify the value to be displayed as the maximum on this axis. If this value is left blank, the maximum will be determined from the data. If this value is less than the largest 90th percentile of the data, it will be ignored.

Major Ticks

Specify the number of large tickmarks and optional grid lines along the axis. A set of minor tickmarks will be generated between each pair of major tickmarks. A reference number is displayed adjacent to each major tickmark.

Minor Ticks

Select the number of small tickmarks to be displayed between each pair of major (large) tickmarks along the axis.

Show Grid Lines

Check this option to display grid lines at the major tickmarks along the axis.

NOTE: Since the grid lines are drawn out from the tickmarks, they appear perpendicular to the corresponding axis. Thus, checking Show Grid Lines here will actually cause horizontal grid lines to appear.

Box Plot Settings

The following options allow you to control the appearance of the box plot.

Style File

Designate a box plot style file. This file sets all box plot options that are not set directly on this panel. Unless you choose otherwise, the Box3 style file is used. Box plot style files are created using the Box Plots procedure.

% Space

When the Box Width (or Bar Width) option is set to Percent Space in the box plot style file selected, this value specifies the percent of the length of the axis that is empty space instead of bars or boxes. It determines the width of the bars or boxes. The smaller this value is, the wider the bars or boxes. Also, note that this parameter only works for non-overlapping bars and boxes.

Whisker

Specify the type of pattern used to display the lines that extend from the edge of the box. The available options are

- **NONE**

  The lines are not displayed.

- **LINE ----**

  The lines are displayed without crossbars.

- **T1 (T-Shape) ----|**

  The lines are displayed as the usual T-shape.

- **T2 (T-Shape with Ticks) ----]**

  The lines are displayed in the usual T-shape with tickmarks facing back toward the box.

Interior

The color used to fill the rectangle formed by the vertical and horizontal axes. Click to change.

Background

The color used behind the plot. Click to change.

Box Fill

The color used to fill the boxes. Click to change.

Box Border

The color used to outline the boxes. Click to change.

Line

Click here to set the properties of the adjacent line such as color, thickness, pattern, etc.

Top and Bottom Titles

Enter text here for the designated title.

REPLACEMENT CODES:

The following codes are replaced by appropriate values when the plot is generated.

{X} is replaced by an appropriate horizontal-axis label.

{Y} is replaced by an appropriate vertical-axis label.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Exporting Expression Data to Multiple Text Files

This section presents an example of how to export three .ges files to separate text files.

The spreadsheet data used are recorded in the GESFILES dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\Utilities* directory.

To run this example, take the following steps or load the **Example 1** template on the ***GESS*** Data Export Engine Template tab).

1 Open the GESFILES dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **GESFILES.S0**.
- Click **Open**.

2 Open the Data Export Engine window.

- From the menus, select **GESS**, then **Data Utilities**, then **Export Data to TXT Files**. The Data Export Engine procedure window will be displayed.
- On the Data Export Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Export Engine window, select the **Variables tab**.
- Under GES File Specifications, set the **GES File Names Variable** to **GES_Files**.
- Under **Folder in which Multiple Text Files** will be Stored, enter **%mydocs_NCSS%\DATA\GESS**.
- Under Multiple Files Storage Options, put a check next to **Overwrite existing output (.txt) files with new output (.txt) files**, **Save GES File Info**, **Save Parent (Input) File Info**, **Save Processing Options**, **Save Files Processed List**, and **Save Parent (Input) File Header**.
- Leave all other options under the Variables tab and other tabs at their default settings.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

GES File Summary

```
GES File Summary

Row     File Name
1       ...\DATA\GESS\Utilities\SampleGES_1.ges
2       ...\DATA\GESS\Utilities\SampleGES_2.ges
3       ...\DATA\GESS\Utilities\SampleGES_3.ges
```

This report displays a list of .ges files exported.

Row

This is the row of the .ges file on the spreadsheet.

File Name

This contains the path and name of each .ges file.

Output File Summary

```
Output File Summary

        Number of
Row     Genes        File Name
1       345          SampleGES_1.txt
2       345          SampleGES_2.txt
3       345          SampleGES_3.txt
```

This report displays a list of the text file names and the number of genes contained in each text file. The folder in which each text file is stored in listed under Output File Specifications.

Row

This is the row of the newly created output file on the spreadsheet. If an input file contains data from more than one array, the output file row may not equal the corresponding input file row.

Number of Genes

This is the number or genes contained in each output file.

File Name

These are the names of the newly created output (.txt) files. The folder into which these files were stored is listed as the "Output File Folder" under Output File Specifications.

Output File Specifications

```
Output File Specifications

Parameter                  Value
GES File Names Variable    GES_Files
Output File Delimiter      Tab
Output File Folder         ...\DATA\GESS
Files Exported             3
```

This report presents the specifications used to create the text files.

Numerical Summary of Expression Values

```
Numerical Summary of Expression Values

                      10th        25th        50th        75th        90th
Row   Minimum   Percentile  Percentile  Percentile  Percentile  Percentile    Maximum
1     3.346565    3.793597    3.965115    4.271953    4.606101    5.031637   6.468214
2     3.320764    3.732144    3.972248    4.274407    4.593267    4.999708   7.798704
3     3.250465    3.752393    4.001728    4.232908    4.610593    5.058788   7.698843
```

This report gives summary statistics of the expression values saved in the output files

Row

This is the row of the .ges file on the spreadsheet.

Minimum

This is the minimum expression value.

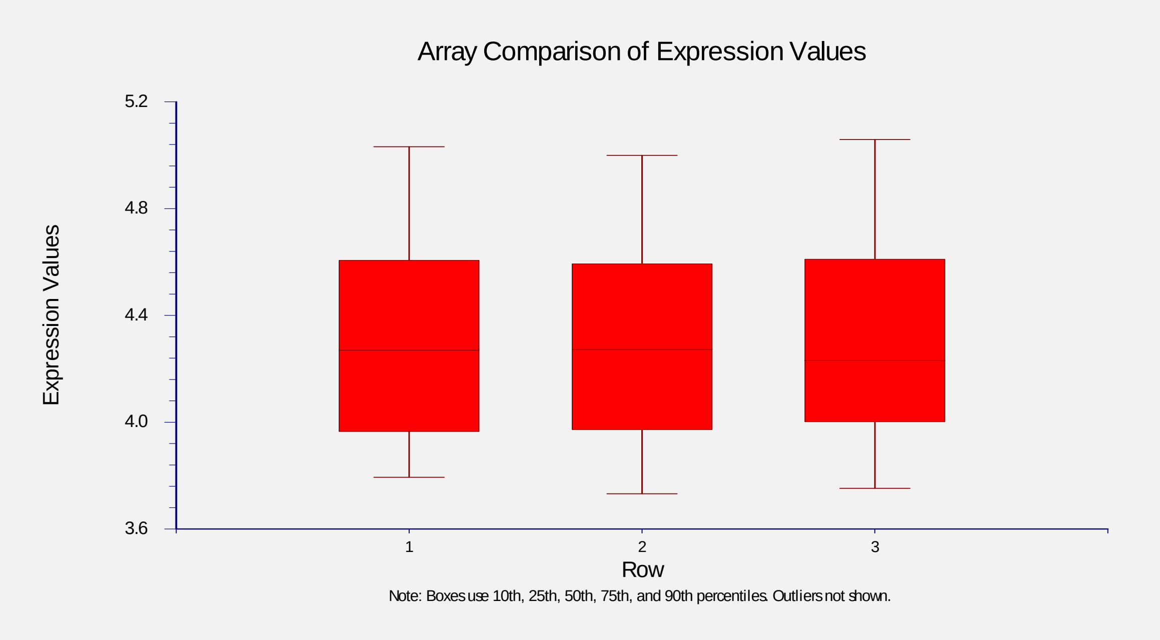
Percentiles

These are the 10th, 25th, 50th, 75th, and 90th expression value percentiles.

Maximum

This is the maximum expression value.

Array Comparison Box Plot



This plot shows the relative distributions of expression values.

Sample Output Text File

The following shows the first several lines of the newly created text file, **SampleGES_1.txt**.

```
<<GES FILE INFO>>
File=...\DATA\GESS\Utilities\SampleGES_1.ges
FileDate=4/12/2006 3:32:12 PM
NumGenes=345
Fingerprint=52812
Database=D:\0A70\DATA\GESS\GEStest.S0
GESFilesVariable=OutputFile

<<PARENT FILE INFO>>
ParentFileType=Affymetrix CEL File
CELFile=D:\0A70\DATA\GESS\SampleGES_1.cel
CDFFile=D:\0A70\DATA\GESS\AF\Test3.cdf
CDHFile=D:\0A70\DATA\GESS\CDH\Test3.cdh
InputVariable=InputFile

<<RMA PROCESSING OPTIONS>>
BackCorr=RMA (Model-Based)
Norm=Quantile Normalization
Summary=Median Polish
OutputScale=Log base 2

<<FILES PROCESSED>>
NumFilesProcessed=6
File1=D:\0A70\DATA\GESS\SampleGES_1.cel
File2=D:\0A70\DATA\GESS\SampleGES_2.cel
File3=D:\0A70\DATA\GESS\SampleGES_3.cel
File4=D:\0A70\DATA\GESS\SampleGES_4.cel
File5=D:\0A70\DATA\GESS\SampleGES_5.cel
File6=D:\0A70\DATA\GESS\SampleGES_6.cel
```

```
<<PARENT FILE HEADER>>
[CEL]
Version=3

[HEADER]
Cols=126
Rows=126
TotalX=126
TotalY=126
OffsetX=0
OffsetY=0
GridCornerUL=154 164
GridCornerUR=995 160
GridCornerLR=999 1003
GridCornerLL=158 1006
Axis-invertX=0
AxisInvertY=0
swapXY=0
DatHeader=[12..40151]  Fetal 3:CLS=1167 RWS=1167 XIN=3  YIN=3  VE=17        2.0 08/16/01
17:28:31    _  _ Test3.1sq _  _  _  _  _  _  _  _  _ 6
Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow:1.004


<<EXPRESSION VALUES>>
Gene                    SampleGES_1.ges
Pae_16SrRNA_s_at        3.99539775521828
Pae_23SrRNA_s_at        4.50946478774071
.                       .
.                       .
.                       .
```

As requested, the file contains a great deal of header information about the .ges file, the parent file, and the processing options. This .ges file exported came from an Affymetrix .cel file that was pre-processed using RMA along with five other .cel files. This information allows for easy tracking of each file. The length and content of the file information depends on the file type.

Example 2 – Exporting Expression Data to a Single Text File

This section presents an example of how to export three .ges files to a single text file.

The spreadsheet data used are recorded in the GESFILES dataset. The input files are stored in the *C:\Program Files\NCSS\NCSS 2007\Data\GESS\Utilities* directory.

To run this example, take the following steps or load the **Example 2** template on the ***GESS*** Data Export Engine Template tab.

1 Open the GESFILES dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **DATA** subdirectory of your **NCSS** directory.
- Open the **GESS** folder.
- Click on the file **GESFILES.S0**.
- Click **Open**.

2 **Open the Data Export Engine window.**

- On the menus, select **GESS**, then **Data Utilities**, then **Export Data to TXT Files**. The Data Export Engine procedure window will be displayed.
- On the Data Export Engine window menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Data Export Engine window, select the **Variables tab**.
- Under GES File Specifications, set the **GES File Names Variable** to **GES_Files**.
- Under TXT Output File Specifications, set **Export Type** to **Single File**.
- Under **Single File Storage Options**, enter **%mydocs_NCSS%\DATA\GESS \ExportSingle.txt** as the **Single Text File Name**.
- Leave all other options under the Variables tab and other tabs at their default settings.

4 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

GES File Summary

```
GES File Summary

Row     File Name
1       ...\DATA\GESS\Utilities\SampleGES_1.ges
2       ...\DATA\GESS\Utilities\SampleGES_2.ges
3       ...\DATA\GESS\Utilities\SampleGES_3.ges
```

This report displays a list of the .ges files exported.

Output File Summary

```
Output File Summary

        Number of
Row     Genes          File Name
1       345            ExportSingle.txt
2       345            ExportSingle.txt
3       345            ExportSingle.txt
*** All GES files were exported to a single output file.
```

This report displays a list of the text file names and the number of genes contained in each text file. Notice that the file name is the same for all rows.

Output File Specifications

```
Output File Specifications

Parameter                 Value
GES File Names Variable   GES_Files
Output File Delimiter     Tab
Output File Folder        ...\DATA\GESS
Files Exported            3
```

This report presents the specifications used to create the text files.

Sample Output Text File

The following shows the first several lines of the newly created text file, ExportSingle.txt.

```
Gene                SampleGES_1.ges     SampleGES_2.ges     SampleGES_3.ges
Pae_16SrRNA_s_at    3.99539775521828    4.65055463471543    5.72425840428734
Pae_23SrRNA_s_at    4.50946478774071    5.16797363869934    3.7614710087377
PA1178_oprH_at      3.71601955717454    5.65006195582932    3.71503950253473
PA1816_dnaQ_at      4.18668328829606    5.21357079221897    5.17736931119034
.                   .                   .                   .
.                   .                   .                   .
.                   .                   .                   .
```

The text file contains the data from all three .ges files. The column titles are the .ges file names.

Chapter 170

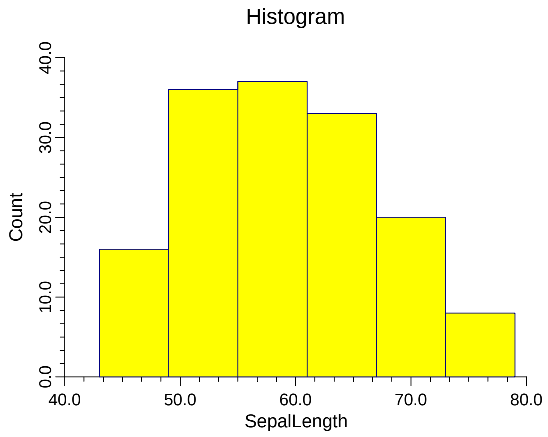
Histograms

Introduction

The word *histogram* comes from the Greek *histos*, meaning pole or mast, and *gram*, which means chart or graph. Hence, the direct definition of "histogram" is "pole chart." Perhaps this word was chosen because a histogram looks like a few poles standing side-by-side.

A histogram is used to display the distribution of data values along the real number line. It competes with the probability plot as a method of assessing normality. Humans cannot comprehend a large batch of observations just by reading them. To interpret the numbers, you must summarize them by sorting, grouping, and averaging. One method of doing this is to construct a *frequency distribution.* This involves dividing up the range of the data into a few (usually equal) intervals. The number of observations falling in each interval is counted. This gives a frequency distribution.

The *histogram* is a graph of the frequency distribution in which the vertical axis represents the count (frequency) and the horizontal axis represents the possible range of the data values.



Density Trace

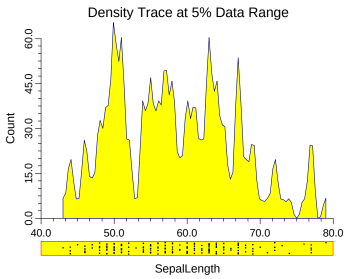
The histogram is widely used and needs little explanation. However, it does have its drawbacks. First, the number and width of the intervals are a subjective decision, yet they have a high impact on the look of the histogram. Slightly different boundary values can give dramatically different looking histograms. (You can experiment with **NCSS** to see the impact of changing the number of bins on the look of the histogram.)

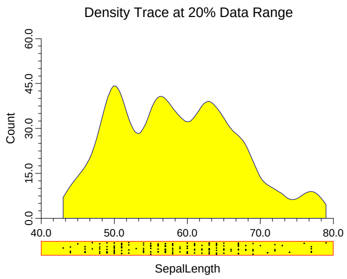
Another problem with the histogram is that the rectangles make it appear that the data are spread uniformly throughout the interval. But this is often not the case. Also, the “skyscraper” look of the histogram doesn’t resemble the rather smooth nature of the data’s distribution.

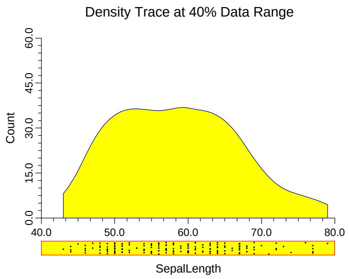
These complaints against the histogram have brought many new innovations. One of the newest and most popular display techniques for showing the distribution of data is the density trace.

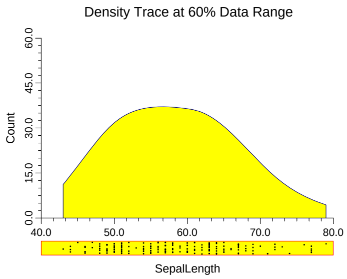
Density refers to the relative frequency (concentration) of data points along the data range. Mathematically, the density at a value x is defined as the fraction of data values per unit of measurement that lie in an interval centered at x. Once you pick a suitable interval width, you can calculate the density at any (and every) x value. If you calculate the density at, say, 50 values and connect them, you’ll have a density trace.

In **NCSS**, the interval width is specified as a percentage. As you increase the percentage, you increase the amount of data included in each density calculation. This increases the smoothness of the chart. The following four density traces were made of the same data at increasing percentage smoothness. Note how much more appealing these charts are than the histogram.









As the interval width is increased, data points further and further from the center value are included. In order to decrease the weight of points that are far removed from the center value, we use a weighting scheme that weights points proportionally to their distance from the center value. The weight function used is half the cosine function with its peak at the center value. It decreases symmetrically to zero, after which a weight of zero is applied. Hence, points have a smaller and smaller impact on the density trace as they are further and further from the center.

Another way to think of the density trace is to imagine that you construct 1000 histograms of the same data using slightly different boundary positions and take the average rectangle height at each of 50 values along the data range. This would give you a smoothed histogram that has many of the same properties of the density trace. Hence, the density trace should be thought of as a smoothed histogram in which interval width and number of bins do not come into play.

Data Structure

A histogram is constructed from a single variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). No other constraints are made on the input data. However, the distributions available in **NCSS** assume that the data are continuous. Note that rows with missing values in one of the selected variables are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the histogram.

Variables

Variable(s)

Select one or more variables. If more than one variable is entered, the values may be combined into one histogram or separated into separate histograms depending on the *Combine all variables as one* option.

Grouping Variable

This variable may be used to separate the observations into groups. A separate histogram is created for each unique value of this variable.

Combine all variables as one

When checked, the values for all selected variables are combined into one histogram. This option cannot be used when a Grouping Variable is specified.

Specify Number of Bars Using

The number of bars shown on the histogram is designated by either the Number of Bars option or the Bar Width option. If the Bar Width option is blank, the Number of Bars is used; otherwise, the Bar Width is used.

Number of Bars

Specifies the number of bars (bins, slices, or intervals) displayed in your histogram. This option is only used if the Bar Width option is blank.

Bar Width

This is the width of the bars in the terms of the data values. It is used in conjunction with the Data Range Minimum and Data Range Maximum to determine the number of bars.

Specify Number of Bars Using – Data Range

Minimum

This is the minimum data value displayed on the histogram. Rows with values smaller than this are omitted. If left blank, it is calculated from the data.

Maximum

This is the maximum data value of the histogram. Rows with values larger than this are omitted. If left blank, it is calculated from the data.

Histogram

Display Type

Indicates whether to fill the bars or overlay only their outline. This option is useful when you want to overlay the histogram outline on a density trace. You can also use this option to completely omit the histogram.

Cumulative Scale

Checking this option causes the program to display the cumulative frequencies instead of the individual frequencies.

Outline

Designates the color, line width, and line pattern of the outline of the histogram's bars.

Histogram – Bar Fill Color

Color

This option specifies the interior color of the histogram bars.

Histogram – Bar Outline

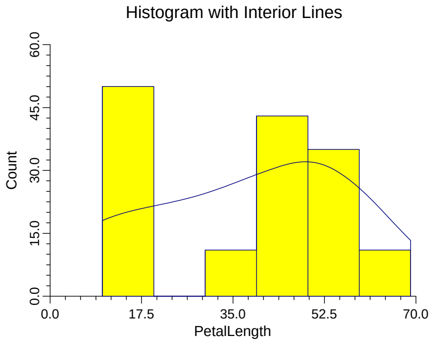
Outline Color, Width, and Pattern

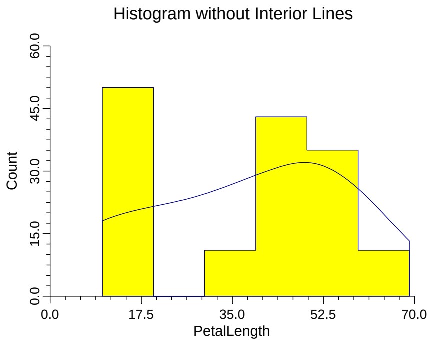
Designates the line color, line width, and line pattern of the outline of the histogram's bars.

Histogram – Interior Lines

Interior Lines

Use this option to omit the internal outline lines that are normally drawn around each bar.





Axes Tab

These options specify the characteristics of the vertical and horizontal axes.

Vertical Axis

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Maximum

Specifies the largest value shown on this axis. Note that the minimum is always set to zero.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Scale Type

Specifies whether the vertical scale is displayed as a Count or a Percentage.

Vertical Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks and grid lines displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

Tick labels are displayed for the minor tickmarks. This option specifies the number of minor tickmarks and grid lines displayed between each set of major tickmarks.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the minor grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the reference numbers that are displayed along this axis. The following options are available in this window:

Decimals

Specifies the number of decimal places displayed in the reference numbers.

Font Size

Specifies the size of the reference values.

Color

Specifies the color of the reference values.

Bold, Italic, Underline

Specifies the font style of the reference values.

Text Rotation

Specifies whether the reference values are displayed vertically or horizontally.

Max Characters

The maximum length (number of characters allowed) of a reference value. This field shifts the axis label away from the axis to make room for the reference value. Hence, if your reference numbers are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

Vertical Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the right side, the left side, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the right side, the left side, on both sides, or not displayed.

Tick Labels

This option controls the position of the tick labels: whether they are placed on the right side, the left side, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Horizontal Axis

These options specifies the characteristics of the horizontal axis.

Label Text

This box supplies the horizontal axis label. The characters {Y} and {G} are replaced by the appropriate variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Location of Bar Labels

This option specifies where the tick labels are placed along the horizontal axis.

Standard

Selecting this option indicates a typical placement of the tick labels. Note that the numbers do not necessarily match the bars.

Mid Points

One tick label is placed at the middle of each bar.

End Points

One tick label is placed at the end of each bar.

Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Color**

  Specifies the color of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Max Characters

The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

Horizontal Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the bottom, the top, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tick Labels

This option controls the position of the tick labels: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Variable Data Transformations

These options specify automatic transformations of the data.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

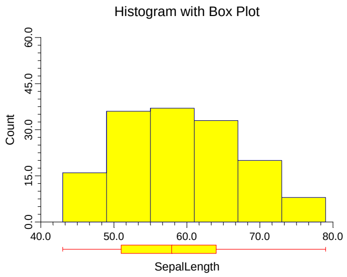
Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Box and Dot Plots Tab

Box Plot

A box plot may be placed above or below the histogram.



Shape

This option specifies the shape of the box plot. See the *Box Plot* chapter for further details.

Percentile Type

Specifies the formula used to calculate the percentile. See the *Box Plot* chapter for further details.

Reference Lines

Reference lines may be extended across the histogram at each of the three quartiles that are used to form the box.

Inner and Outer Fences

The constants used to construct the fences. See the *Box Plot* chapter for further details.

Position

This option indicates on which side of the histogram the box plot should be displayed.

Fill Color

This option specifies the interior color of the box plot.

Line Color

This option specifies the color of the box border and the lines.

Line Width

This option specifies the width of the border line.

Box Width

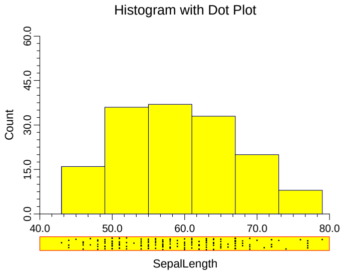
This option specifies the width of the box itself.

Show Outliers

This option indicates whether the outlying points should be displayed.

Dot Plot

A dot plot may be placed in the vertical margin of the plot. The dot plot lets you study the distribution of the data values.



Position

Specifies whether to display the dot plot and in which margin to place it in.

Box Width

Specifies the width of the dot plot.

Fill Color

The color of the data point interior (fill region). Many of the plotting symbols, such as a circle or square, have an interior region and a border. This specifies the color of the interior region.

Dot Color and Size

Specifies the color and size of the data points displayed in the dot plot.

Line Width and Color

Specifies the width and color of the dot plot’s border.

Traces and Lines Tab

These options allow certain reference lines to be specified and displayed.

Trace and Line Overlays – Density Trace

This set of options controls the appearance of the optional density trace. A density trace may be placed on, or replace, the histogram. The details of how a density trace is constructed were presented at the beginning of this chapter.

Display Type

Indicates whether to fill the density trace or overlay only its outline. This option is useful when you want to overlay the density trace outline on a histogram. You can also use this option to completely omit the density trace.

Number of Points

Specifies the number of density trace points to be calculated along the horizontal axis. This adjusts the resolution of the density trace.

Percent of Data in Calculation

The percent of the data used in the density calculation. Select 0 for an automatic value determined from your data. A low value (near 10%) will give a rough plot. A high value (near 40%) will yield a smooth plot.

Outline Color, Width, and Pattern

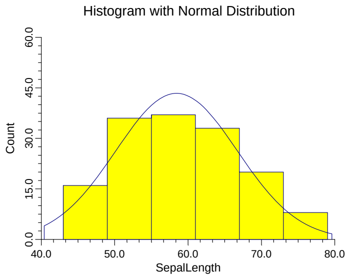
Designates the color, line width, and line pattern of the outline.

Fill Color

This option specifies the interior color of the density trace.

Trace and Line Overlays – Normal Line

This panel controls the appearance of a normal density line. This line uses the estimated mean and standard deviation to overlay a normal density function.



Display Type

Indicate whether to fill the normal density or overlay only its outline. This option is useful when you want to overlay the normal density outline on a histogram or density trace. You can also use this option to completely omit the normal density line.

Number of Calculation Points

Specifies the number of points to be calculated along the horizontal axis. This adjusts the resolution of the normal density curve.

Outline Color, Width, and Pattern

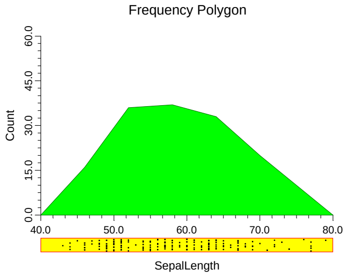
Designates the color, line width, and line pattern of the outline.

Fill Color

This option specifies the interior color of the normal density.

Trace and Line Overlays – Frequency Line

These options control the frequency polygon that may replace—or be added to—the histogram. The frequency polygon is a line that connects the top midpoints of the histogram's bars.



Display Type

Indicate whether to fill the frequency polygon or overlay only its outline. This option is useful when you want to overlay the frequency polygon outline on a histogram. You can also use this option to completely omit the frequency polygon. Note that the number of intervals (bins) is set in the Histogram Tab section.

Outline Color, Width, and Pattern

Designates the color, line width, and line pattern of the outline of the frequency polygon.

Fill Color

This option specifies the interior color of the frequency polygon.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

This option lets you display up to two horizontal lines at particular values. The actual value is specified to the right of the line.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

This option lets you display up to two vertical lines at particular values. The actual value is specified to the right of the line.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Histogram

This section presents an example of how to generate a histogram. The data used are from the FISHER database. We will create a histogram of *SepalLength*.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Histograms window.

1 **Open the FISHER dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 **Open the Histograms window.**

- On the menus, select **Graphics**, then **Histograms**. The Histograms procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Histograms window, select the **Variables tab**.
- Double-click in the **Data Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. “SepalLength” will appear in the Data Variable(s) box.

4 **Specify the title.**

- On the Histograms window, select the **Titles and Misc. tab**.
- In the **Top Title Line 1** box, enter **Histogram With Everything On It.**

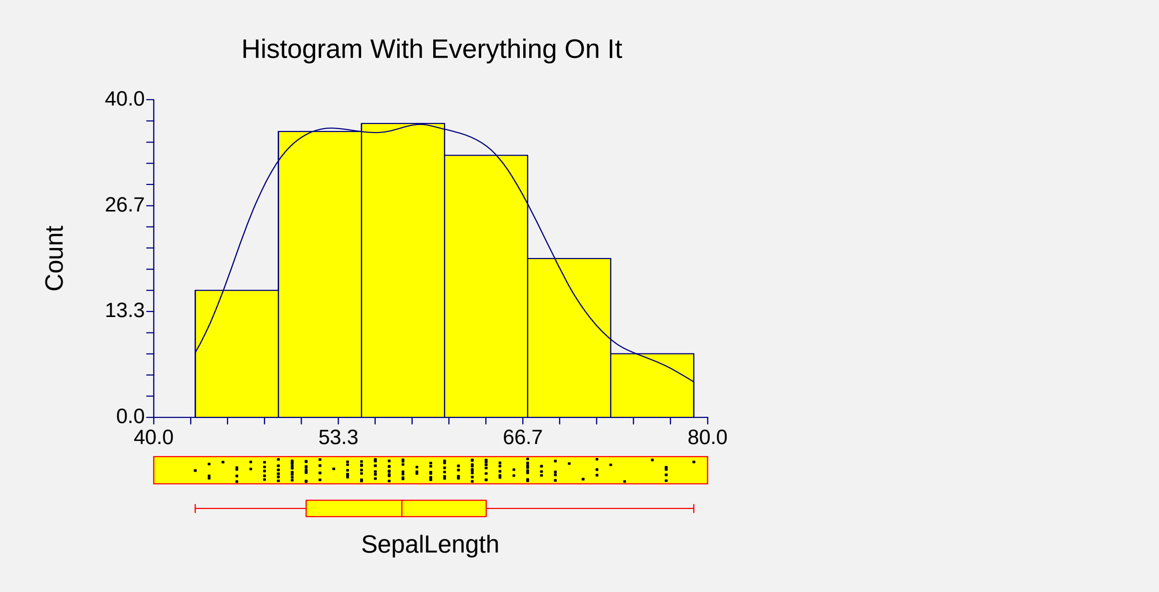
5 **Specify the Box Plot and the Dot Plot.**

- On the Histograms window, select the **Box and Dot Plots tab**.
- In the **Shape box under Box Plot**, select **Rectangle**.
- In the **Position box under Box Plot**, select **Bottom**.
- In the **Position box under Dot Plot**, select **Bottom**.

6 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Histogram Output



This histogram has a density trace overlay with a dot plot and box plot below it. This chart allows you to study almost all univariate features of this variable!

Creating a Histogram Style File

Many of the statistical procedures include histograms as part of their reports. Since the histogram has almost 200 options, adding it to another procedure's report greatly increases the number of options that you have to specify for that procedure. To overcome this, we let you create and save histogram style files. These files contain the current settings of all histogram options. When you use the style file in another procedure you only have to set a few of the options. Most of the options come from this style file. A default histogram style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a histogram style file.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.
- Note: You do not necessarily have to use the FISHER database. You can use whatever database is easiest for you. Just open a database with a column of numeric data.

2 Open the Histograms window.

- On the menus, select **Graphics**, then **Histograms**. The Histograms procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Histograms window, select the **Variables tab**.
- Double-click in the **Data Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Data Variable(s) box.
- Note: don't worry about the specification of this variable. The actual variable names are not stored in the style file. They are used here so that you can see what your style will look like.

4 Set your options.

- Set the various options of the histogram's appearance to the way you want them.
- Run the procedure to generate the histogram. This gives you a final check on whether it appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 Save the template (optional).

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template file is not the style file.
- To store the template, select the **Template tab** on the Histogram window.
- Enter an appropriate name in the **File Name** box.
- Enter an appropriate phrase at the bottom of the window in the **Template Id** (the long box across the bottom of the Histograms window). This phrase will be displayed in the Template Id's box to help you identify the template files.
- Select **Save Template** from the File menu. This will save the template.

6 Create and Save the Style File

- Select **Save Style File** from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the **Selected File** box. You can either reuse one of the style files that already exist or create a new name. You don't have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 Using a Style File

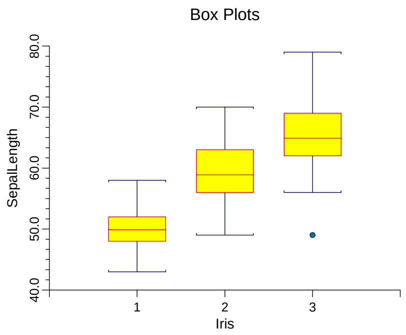
- Using the style file is easy. For example, suppose you want to use this style file to plot residuals in the Multiple Regression procedure. You do the following:
- Select the **Plot Options tab** in the Multiple Regression procedure.
- Click the **button to the right of the Histogram - Plot Style File box** (the initial file name is Default). This will bring up the Histogram Style File Selection window.
- Click on the appropriate file so that it is listed in the **Selected File** box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Multiple Regression window. That's it. Your new style has been activated.

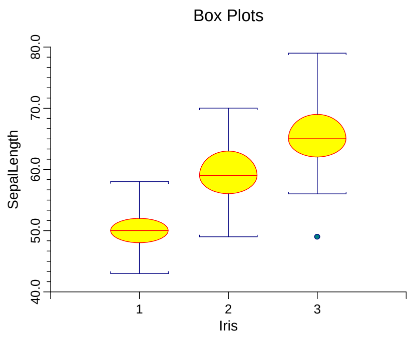
Chapter 171

Box Plots

Introduction

When analyzing data, you often need to study the characteristics of a single batch of numbers, observations, or measurements. You might want to know the center and how spread out the data are about this central value. You might want to investigate extreme values (referred to as outliers) or study the distribution of the data values (the pattern of the data values along the measurement axis). Several techniques are available to allow you to study the distribution. These include the stem-leaf plot, histogram, density trace, probability plot, and box plot.





Box Plot Definition

The box plot shows three main features about a variable: its center, its spread, and its outliers.

Box

A box plot is made up of a box (a rectangle) with various lines and points added to it. The width of the box is arbitrary and should be selected to make an eye-pleasing display. The top and bottom of the box are the 25th and 75th percentiles. The length of the box is thus the interquartile range (IQR). That is, the box represents the middle 50% of the data. The IQR is a popular measure of spread. You can represent the box as a rectangle, a diamond, an ellipse, or a special figure designed for making multiple comparisons.

A line is drawn through the middle of the box at the median (the 50th percentile). The median is a popular measure of the variable’s location (center or average value).

Adjacent Values

The *upper adjacent value* is the largest observation that is less than or equal to the 75th percentile plus 1.5 times IQR. The *lower adjacent value* is the smallest observation that is greater than or equal to the 25th percentile minus 1.5 times IQR.

The adjacent values are displayed as T-shaped lines that extend from each end of the box.

Outside Values

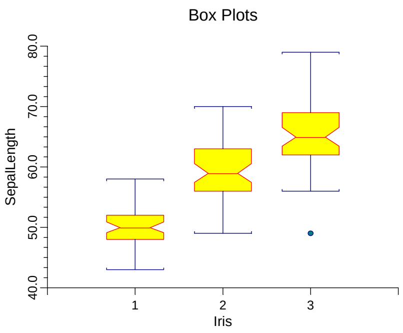
Values outside the upper and lower adjacent values are called *outside values*. Values that are under three IQRs from the 25th and 75th percentiles are called *mild* outliers. Those outside three IQRs are called *severe* outliers. Mild outliers are not unusual, but severe outliers are.

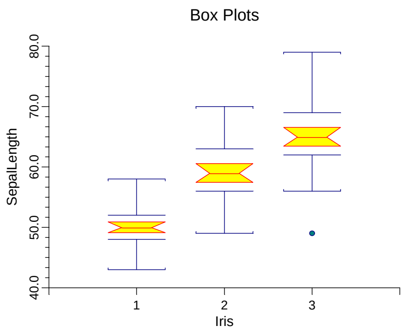
Multiple Comparisons

Box plots are often used for comparing the distributions of several batches of data, since they summarize the center and spread of the data very nicely. When making strict comparisons among the locations (medians) of various batches, a modified box plot called the *notched box plot* is useful. The notches are constructed using the formula:

$$Median \pm 1.57 \times (IQR)/\sqrt{n}$$

Notched box plots are used to make multiple comparisons among the batches. If the notches of two boxes do not overlap, we may assume that the medians are significantly different (the centers are statistically significant). The 1.57 is selected for the 95% level of significance. The box plot on the left is the classical notched box plot.





Recently, statisticians have noticed that the notched box plot does not allow you to focus on the multiple comparisons. A modern version of the notched box plot has been proposed that lets you make this comparison (see the above plot on the right). This version modifies the symbol used for the box. In fact, it leaves the box out. Two horizontal lines mark the position of the box. The part that is plotted is the notched part only. This makes it much easier to make comparisons. If two of the notches overlap, the group medians are not significantly different. Otherwise, they are.

Note that when making comparisons among several batches, the notched box plots do not make any adjustment for the multiplicity of tests being conducted. As long as the notched box plots are used informally, no technical adjustments are necessary.

Data Structure

A box plot is constructed from one variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). In this case, a separate box plot is displayed for each group. No other constraints are made on the input data.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the box plot.

Variables

Variable(s)

This option lets you designate which variables are plotted. If more than one variable is designated and no Grouping Variable is selected, a set of box plots will be displayed on a single chart, one box for each variable. If more than one variable is designated and a Grouping Variable is selected, a separate box plot will be drawn for each variable.

Grouping Variable

Designates an optional variable used to separate the observations into groups. An individual box will be displayed for each unique value of this variable.

Data Label Variable

A data label is text that is displayed beside each outside point. This option designates the variable containing the data labels. The values may be text or numeric.

Box

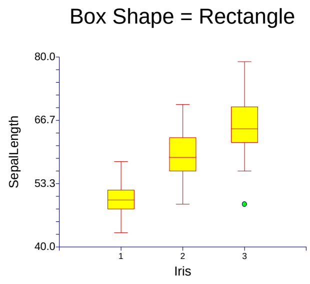
Shape

This option specifies the type (shape) of the box plot. Possible shapes are rectangle (standard), diamond, ellipse, and notched (original and modified). The special shapes are used for making comparisons among several groups (see Multiple Comparisons earlier in this chapter).

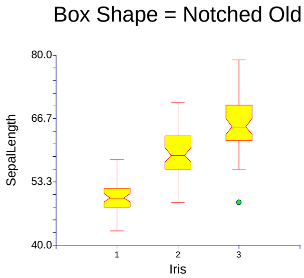
- **None**

  The box plot is not displayed.

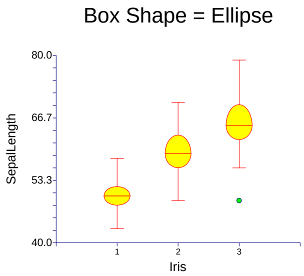
- **Rectangle**



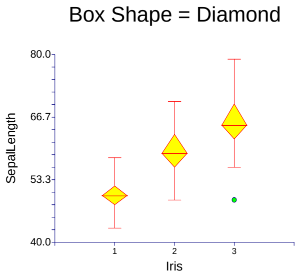
- **Notched – Old**



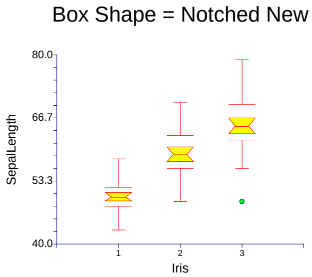
- **Ellipse**



- **Diamond**



- **Notched – New**



Median Symbol

Click in the box or on the button to its right to display a window that allows you to change the characteristics of the median symbol.

Note: If a horizontal line is selected (which is the default), it will be ignored and the outline color and width will be used instead.

Percentile Calculation Method

Specify the formula used to calculate the percentile. See the Descriptive Statistics chapter for further details.

Box – Outline

Color

This option specifies the color of the box border and the lines.

Width

This option specifies the width of the border line.

Box – Fill

Color

This option specifies the interior color of the boxes.

Box – Width

Box Width Parameter

This option designates how you want to specify the width of boxes. You can specify an actual Amount or a Percent Space.

Amount

Specify the exact width of the box. It is only used if it is specified in Use for Box Width above.

Percent Empty Space

This option specifies what percent of the horizontal axis should be kept as “white space.” The smaller this value is, the larger the box width will be. It is only used if it is specified in Use for Box Width above.

Lines (Whiskers)

The *adjacent value* is the line that extends up and down from the box. The following options concern the display of these lines.

Type

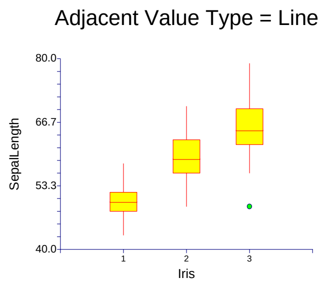
Specifies the type of pattern used to display the adjacent values. Possibilities include a simple line, a T-shaped line, a T-shaped line with backward extenders, and none (omit the adjacent values).

comparisons among several groups (see Multiple Comparisons earlier in this chapter).

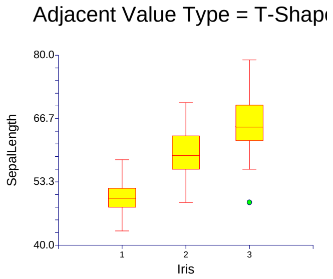
- **None**

  The adjacent values (lines) are not displayed.

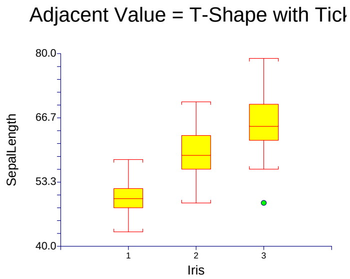
- **Line**



- **T-Shape**



- **T-Shape with Ticks**



Line

This option specifies the color, width, and pattern of the adjacent value line.

Lines (Whiskers) – Type = T-Shape with Ticks

Tick Length

The length of the T-shaped with Ticks option's backward extenders.

Outliers

Show Outliers

This option indicates whether the outlying points should be displayed. It also lets you designate whether you want to show only severe outliers or all outside points.

Outliers - Mild / Severe

Fence Multiplier

These are the constant used to construct the fences. The adjacent values are the *inner fences*. The boundary for declaring an outside value as mild or severe is the outer fence. These values are typically set at 1.5 and 3.0, respectively. These options let you manipulate these constants.

Symbol

This option designates the appearance of the symbol used to portray mild and severe outliers. Click the button on the right of the box to change features of the symbol such as the color and type.

Axes Tab

These options specify the characteristics of the vertical and horizontal axes.

Vertical Axis

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical Axis – Tickmarks and Grid Lines

Major Ticks (number)

Reference numbers are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the right side, the left side, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the right side, the left side, on both sides, or not displayed.

Tick Labels

This option controls the position of the reference numbers: whether they are placed on the right side, the left side, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Horizontal Axis

These options specifies the characteristics of the horizontal axis.

Label Text

This box supplies the horizontal axis label. The characters {Y} and {G} are replaced by the appropriate variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Horizontal Axis – Tickmarks

Ticks

This option sets the color, length, width, and pattern of the tickmarks along this axis. The tickmarks are positioned one per individual box plot.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Horizontal Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the bottom, the top, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tick Labels

This option controls the position of the reference numbers: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls these titles. In the text, the characters *{Y}* and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls these titles. In the text, the characters *{Y}* and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

Background Color

The background color of the plot.

Interior Color

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display a variable’s name, its label, or both the name and label.

Value Labels

This option selects whether to display values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Variable Data Transformation

Transform Exponent

Each value of the variable is raised to this exponent automatically before it is processed. Fractional exponents require positive data values.

Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Lines Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

This option lets you display a horizontal line at a particular value. The actual value is specified to the right of the line.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

This option lets you display a vertical line at a particular value. The actual value is specified to the right of the line.

Horizontal Reference Lines at Median and Box Ends

Reference Line Type

Reference lines may be extended across the plot at each of the three quartiles that are used to form the box.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Box Plot

This section presents an example of how to generate a box plot. The data used are from the FISHER database. We will create box plots of the *SepalLength* variable, breaking on the type of iris.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Box Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Box Plots window.

- On the menus, select **Graphics**, then **Box Plots**. The Box Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

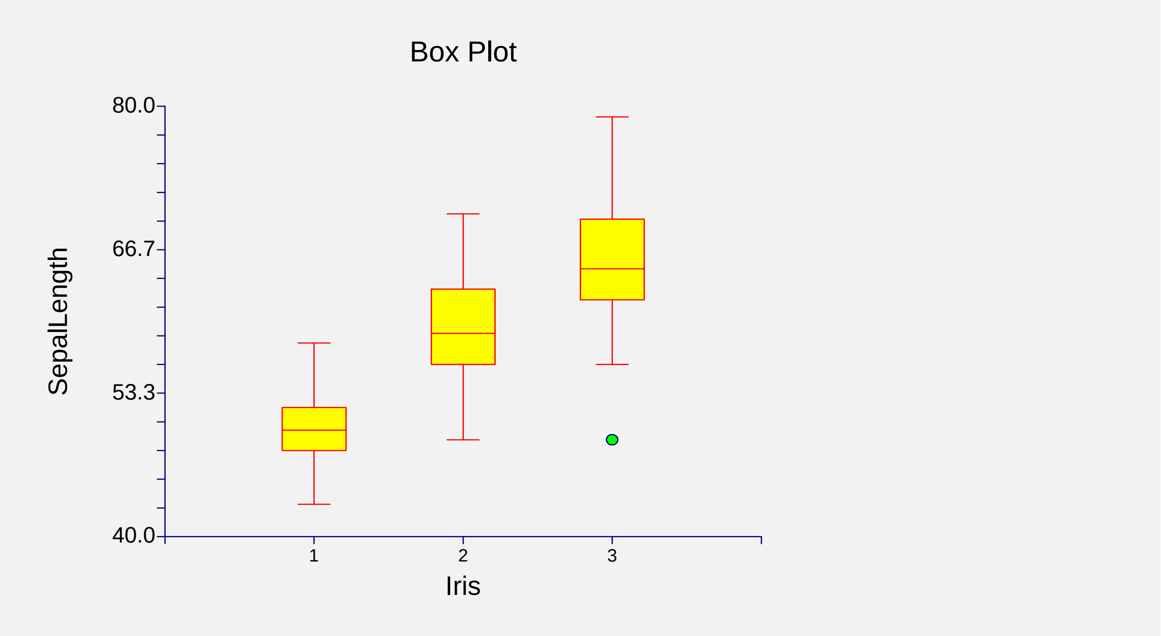
3 Specify the variables.

- On the Box Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. “SepalLength” will appear in the Variable(s) box.
- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Grouping Variable box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Box Plot Output



Creating a Box Plot Style File

Many of the statistical procedures include box plots as part of their reports. Since the box plot has almost 200 options, adding it to another procedure's report greatly increases the number of options that you have to specify for that procedure. To overcome this, we let you create and save box plot style files. These files contain the current settings of all box plot options. When you use the style file in another procedure you only have to set a few of the options. Most of the options come from this style file. A default box plot style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a box plot style file.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.
- Note: You do not necessarily have to use the FISHER database. You can use whatever database is easiest for you. Just open a database with a column of numeric data.

2 Open the Box Plots window.

- On the menus, select **Graphics**, then **Box Plots**. The Box Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Box Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. “SepalLength” will appear in the Variable(s) box.
- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Grouping Variable box.

4 Set your options.

- Set the various options of the box plot’s appearance to the way you want them.
- Run the procedure to generate the box plot. This gives you a final check on whether it appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 Save the template (optional).

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template file is not the style file.
- To store the template, select the **Template tab** on the Box Plot window.
- Enter an appropriate name in the File Name box.
- Enter an appropriate phrase at the bottom of the window in the Template Id (the long box across the bottom of the Box Plot’s window). This phrase will be displayed in the Template Id’s box to help you identify the template files.
- Select **Save Template** from the File menu. This will save the template.

6 Create and Save the Style File

- Select **Save Style File** from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the **Selected File** box. You can either reuse one of the style files that already exist or create a new name. You don’t have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 Using a Style File

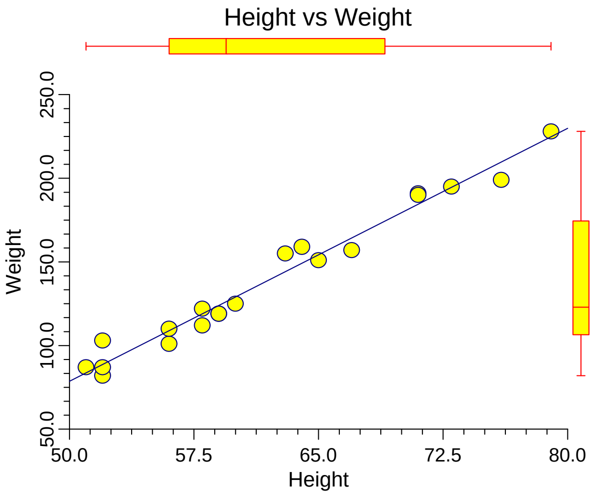
- Using the style file is easy. For example, suppose you want to use this Box Plot Style file in the Two-Sample T-Test procedure. You do the following:
- Select the **Box Plot tab** in the Two-Sample T-Test procedure.
- Click the button to the right of the **Plot Style File** box (the initial file name is Default). This will bring up the Box Plot Style File Selection window.
- Click on the appropriate file so that it is listed in the **Selected File** box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Two-Sample T-Test window. That’s it. Your new style has been activated.

Chapter 172

Scatter Plots

Introduction

The x-y scatter plot is one of the most powerful tools for analyzing data. **NCSS** includes a host of features to enhance the basic scatter plot. Some of these features are trend lines (least squares) and confidence limits, polynomials, splines, lowess curves, imbedded box plots, and sunflower plots. Following is an example of a typical scatter plot with a trend line and imbedded box plots.



Data Structure

A scatter plot is constructed from two variables. A third variable may be used to divide the first two variables into groups (e.g., age group or gender). No other constraints are made on the input data. Note that rows with missing values in one of the selected variables are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are in the scatter plot.

Variables

Vertical Variable(s)

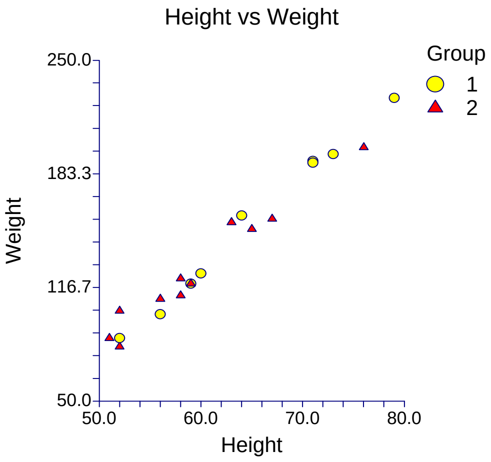
Enter one or more vertical variables. If more than one variable is entered, the number of plots is determined by the *Overlay* option.

Horizontal Variable(s)

Enter one or more horizontal variables. If more than one variable is entered, the number of plots is determined by the *Overlay* option.

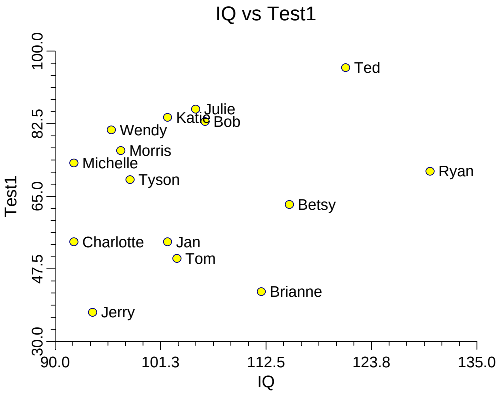
Grouping (Symbol) Variable

This variable may be used to separate the observations into groups. For example, you might want to use different plotting symbols to distinguish observations from different states. You designate the grouping variable here. The appearance of the plot symbol is designated on the Symbols Tab. Labels may be automatically substituted for values if the Value Labels option is set to Value Labels.



Data Label Variable

A data label is text that is displayed beside each point. A variable containing the data labels. The values may be text or numeric. You can use dates (like Jan-23-95) as labels. Here is how. First, enter your dates using the standard date format (like 06/20/93). In the Variable Info screen, change the format of the date variable to something like *mmm-dd-yyyy* or *mm-dd-yy*. The labels will be displayed as labels. Without changing the variable format, the dates will be displayed as long integer values.



The size, style, and color of the text may be modified by pressing the second button to the right of the text box. This button brings up the text settings window.

Plot Overlay

This option is used when multiple vertical and/or horizontal variables are entered to specify whether to overlay specified plots onto a single plot. Possible choices are:

- **None**

  No plots are overlaid. Each combination of horizontal and vertical plots produces its own separate plot.

- **Multiple Vertical**

  All vertical variables are combined onto one plot. A separate plot is drawn for each horizontal variable.

- **Multiple Horizontal**

  All horizontal variables are combined onto one plot. A separate plot is drawn for each vertical variable.

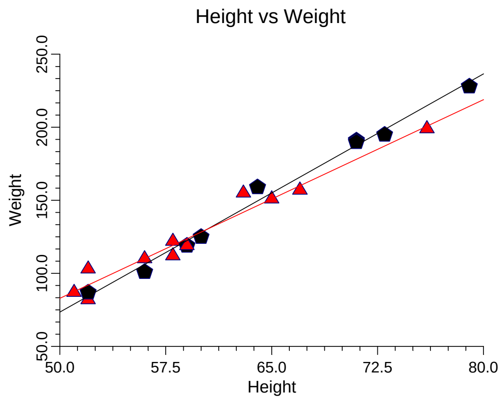
- **Series: No Overlay**

  A series of plots are drawn using each vertical and horizontal variable in sequence. The first vertical variable is matched with the first horizontal variable, the second vertical with the second horizontal, and so on.

Symbols

These options set the type, color, size, and style of the plotting symbols. Symbols for up to fifteen groups may be used. When no Group Variable is specified, the options of Symbol 1 are used to define the plot symbol.

Following is an example of possible symbol settings for two groups:



Double-clicking the symbol, or clicking the button to the right of the symbol, brings up the symbol specification window. This window lets you specify the characteristics of a symbol in detail.

- **Symbol Fill Color**

  The color of the symbol's interior (fill region). Many of the plotting symbols, such as a circle or square, have both an interior region and a border. This specifies the color of the interior region. The border is specified as the Symbol Outline.

- **Symbol Outline Color**

  The color of the symbol's border. This color is used when the symbols are connected by a connecting line.

- **Symbol Type**

  This option designates the shape of the plot symbol. The most popular symbols may be designated by pressing the appropriate button. About 80 symbol types are available, including letters and numbers.

- **Symbol Radius**

  The size (radius) of the symbol.

- **Symbol Fill Pattern**

  The pattern (solid, lines, etc.) of the symbol's interior (fill region).

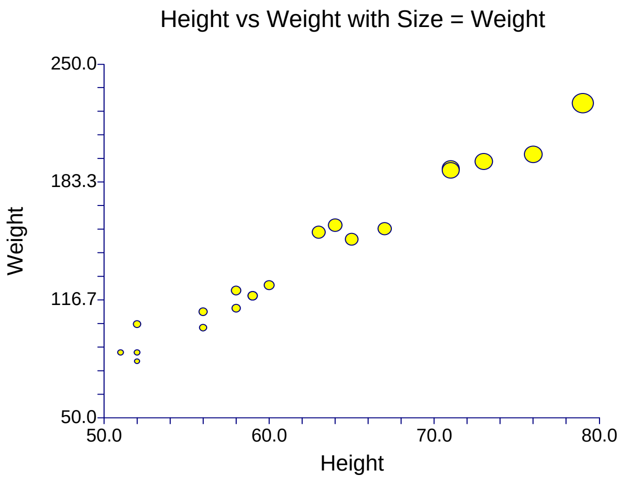
- **Symbol Border Width**

  The width of the symbol's border.

Symbols – Symbol Size Options

Symbol Size Variable

This option designates a third variable that is represented by the size the plotting symbol. Sometimes, this is referred to as a bubble chart. Here is an example of such a plot:



The minimum and maximum values of this variable are associated with the smallest and largest plot symbols. The rest of the points fall in between. The size of the smallest and largest points is controlled by the Minimum and Maximum options explained next.

Minimum Symbol Size

This is the size of the smallest plotting symbol—the one that represents the minimum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 50 means that the diameter of the plotting symbol is one half of normal. Note that normal is represented by 100.

Maximum Symbol Size

This is the size of the largest plotting symbol—the one that represents the maximum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 200 means that the diameter of the plotting symbol is twice that of normal. Note that normal is represented by 100.

Axes Tab

These options specify the characteristics of the vertical and horizontal axes.

Vertical and Horizontal Axis

Label Text

This box supplies the axis label. The characters {X}, {Y}, and {G} are replaced by the horizontal, vertical, and grouping variable names, respectively. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line and its type (numeric or text).

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the scatter plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. The characters {A} and {B} are replaced by the numeric values of the intercept and slope of the regression line, respectively. To display the fitted regression equation, you could use {Y} = {A} + ({B}){X}.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable’s name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Legend

When data for more than one group are displayed, a legend is desirable. These options specify the legend.

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. The characters *{G}* will be replaced by the name of the group variable. Click the button on the right to specify the font size, color, and style of the legend text.

Variable Data Transformations

These options specify automatic transformations of the data for either variable.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

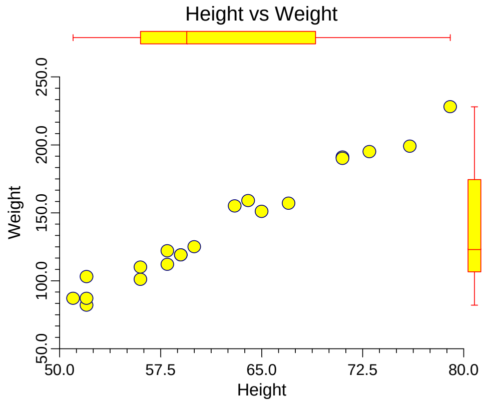
Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Box & Dot Plot Tab

Box Plots

Box plots may be placed in the vertical and horizontal margins of the scatter plot. These box plots emphasize the univariate behavior of the corresponding variable.



Shape

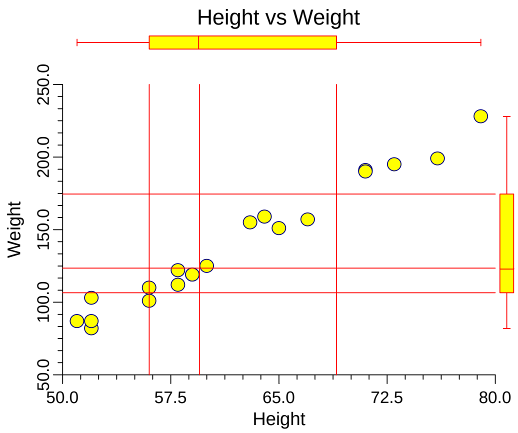
This option specifies the shape of the box plot. See the *Box Plot* chapter for further details.

Percentile Type

Specifies the formula used to calculate the percentile. See the *Box Plot* chapter for further details.

Reference Lines

Reference lines may be extended across the plot at each of the three quartiles that are used to form the box.



Inner and Outer Fences

These multipliers are used to construct the fences for designating outliers. See the *Box Plot* chapter for further details.

Line Width

This option specifies the width of the border line.

Box Width

This option specifies the width of the box itself.

Position

This option indicates on which side of the plot the box plot should be displayed.

Show Outliers

This option indicates whether the outlying points should be displayed.

Fill Color

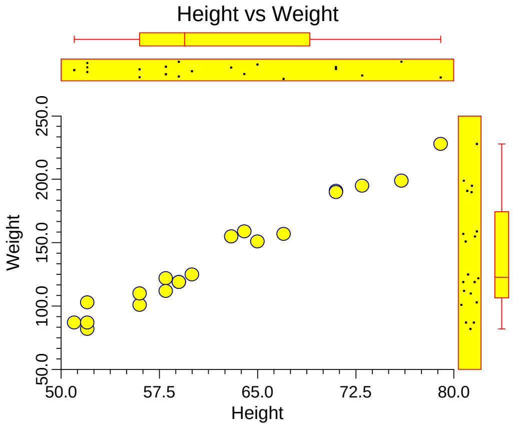
This option specifies the interior color of the box plot.

Line Color

This option specifies the color of the box border and the lines.

Dot Plots

A dot plot may be placed in the vertical and horizontal margins of the scatter plot. The dot plot lets you study the distribution of the corresponding variable by plotting the actual data values in a line plot.



Position

Specifies whether to display the dot plot and in which margin to place it in.

Box Width

Specifies the width of the dot plot.

Dot Color and Size

Specifies the color and size of the data points displayed in the dot plot.

Line Width and Color

Specifies the width and color of the dot plot’s border.

Fill Color

The color of the data point interior (fill region). Many of the plotting symbols, such as a circle or square, have an interior region and a border. This specifies the color of the interior region.

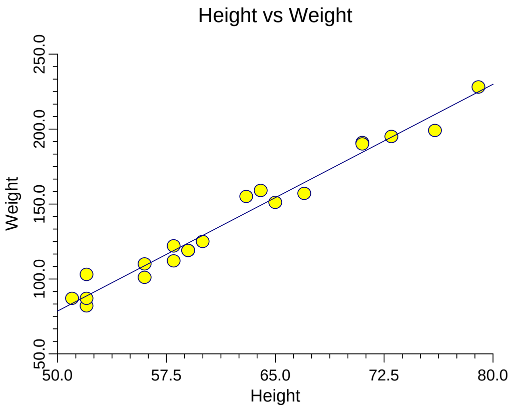
Lines 1 Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line. Several of the lines may be further defined using the pull-down box to the left of the line.

Regression

Regression Line

This option governs the display of a regression line (least-squares trend line or line of best fit) through the data points.



Regression Estimation

This option specifies the way in which the trend line is calculated.

- **L.S.**

  The standard least squares regression line is calculated. This formulation is popular but can suffer from severe distortion if one or more outliers exist in your data.

- **Median**

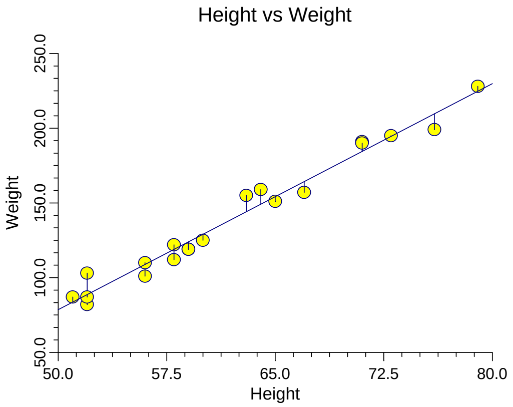
  A resistant least-squares algorithm is used to fit the line, and the intercept is adjusted so that the line passes through the medians of the horizontal and vertical variables. This algorithm is resistant because it removes most of the distortion caused by a few outliers (atypical points).

- **Quartiles**

  A resistant least-squares algorithm is used to fit the line, and the intercept is adjusted so that the line passes through the first and third quartiles of the horizontal and vertical variables. This algorithm is resistant because it removes most of the distortion caused by a few outliers (atypical points).

Residual Lines

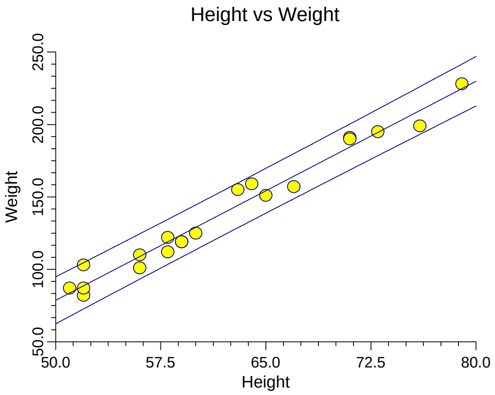
The residual is the vertical deviation of each point from the regression line. These residuals may be displayed as vertical lines that connect each plot point to the regression line.



Confidence Limits

C.L. Individuals (or Means)

Confidence limits for the regression line may be displayed. These limits are for either the mean of a future group of individuals with a common horizontal value or for a single individual. The most commonly used confidence interval is for a future individual. The confidence limits for the mean are for specialized use only. Note that these confidence lines are formed by calculating individual confidence limits at various points along the horizontal axis and connecting those points. These are not the same as confidence bands.



Confidence Limit Alpha

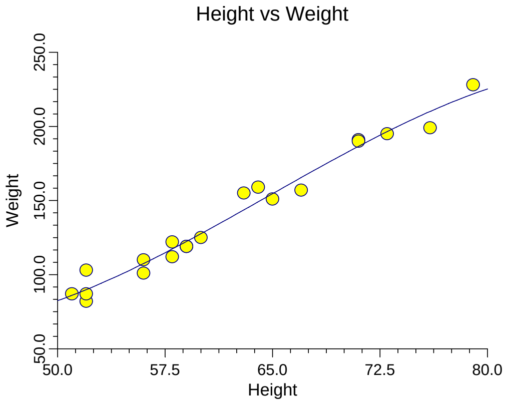
This option gives the confidence level, α, of the 100(1-α)% confidence limits. For example, if α is set to 0.05, you have a 95% confidence limit constructed at each point.

Polynomial Fit

Polynomial Fit

An n-degree polynomial line may be fit and displayed over the data. The number of points along the line that are calculated is controlled by the Number of Calculation Points setting.

If you wish to see the equation of this line, use the response surface regression procedure.



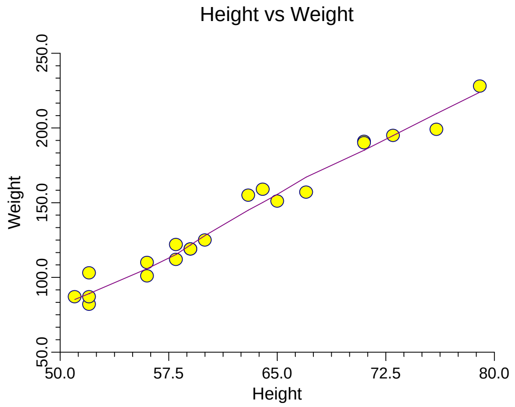
Polynomial Order

The order of the polynomial is the value of the largest exponent in the polynomial regression equation. Usually, values of two or three are used, since polynomials of larger order often exhibit strange behavior between data points.

LOESS

Loess Line

The locally weighted regression scatter plot smooth (*lowess* or *loess*) is a popular, computer-intensive technique that usually provides a reasonable smoothing of your data without being overly sensitive to outliers. A reasonable smooth is one that travels more or less through the middle of the data. The degree of smoothing is controlled by the loess options. The number of points at which the loess curve is computed is given by the *Intervals* option. The mathematical details of the loess method are given in the Linear Regression chapter.



Calculation Fraction

This is the percent of the sample points that are included in the computation for a particular value of the smooth. Most authors recommend 40% as the first value to try. This means that 40% of the data values are used in the computations for each loess smooth value.

LOESS Polynomial Order

This is the order of the polynomial fit in the loess procedure. Select ‘1’ for a linear fit or ‘2’ for a quadratic fit. The linear fit tends to be smoother. The quadratic fit tends to pick up peaks and values better.

Robust Iterations

This is the number of robust iterations used in the loess algorithm to downplay the influence of outliers. Select ‘0’ if you do not want robust iterations. This will show the impact of outliers on the loess curve. It will reduce the execution time in large datasets.

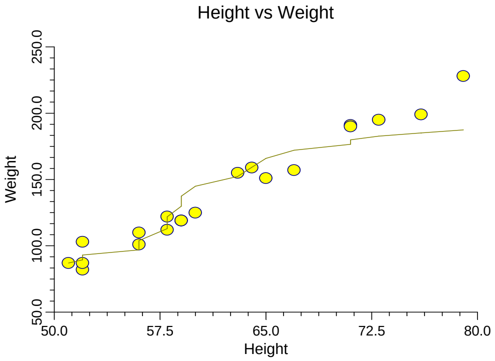
There is little reason for using more than two iterations.

Median Smooth

Median-Smooth Line

A median smooth line may be displayed over the data. This type of smooth line does well for level series (those with no vertical trend), but does not do well when a trend is apparent.

The median smooth is constructed by first ordering the observations by the horizontal variable. Next, running medians of *n* observations are found, where *n* is the Rows parameter.



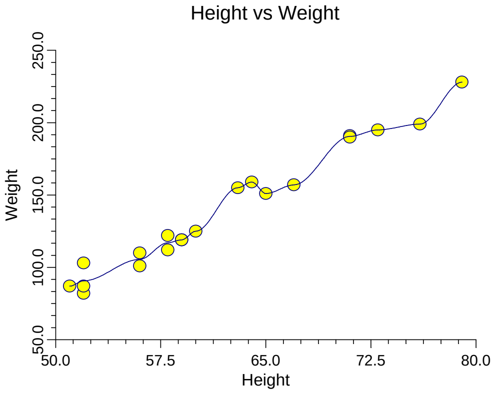
Rows in Smooth Calculation

The number of observations used to calculate the median at each step.

Spline

Spline Fit

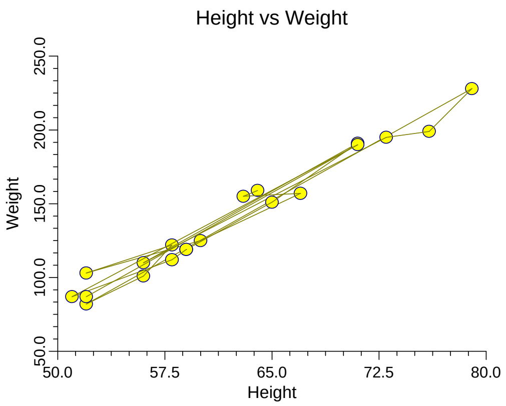
A cubic spline may be fit to the data.



Connect Points

Connect All Points

The points may be connected with a line sequentially, proceeding from the first row, to the second row, to the third row, and so on.



Calculation Points

Number of Calculation Points

The number of positions along the horizontal axis at which the confidence limits, LOESS, and splines are calculated.

Lines 2 Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Horizontal Lines from the Vertical Axis

Line at Mean of Vertical Variable

Display a line at the mean value of the vertical (horizontal) variable.

Line from Vertical Axis to Data Points

This option lets you display individual lines from the data points to the vertical (or horizontal) axis.

Horizontal Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Vertical Lines from the Horizontal Axis

Line at Mean of Horizontal Variable

Display a line at the mean value of the vertical (horizontal) variable.

Line from Horizontal Axis to Data Points

This option lets you display individual lines from the data points to the vertical (or horizontal) axis.

Vertical Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Axis Tickmarks for Data Points

Vertical Axis Ticks for each Data Point

Display tick marks for each data point along the vertical axis.

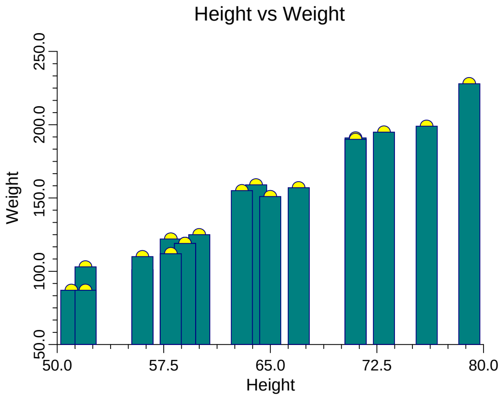
Horizontal Axis Ticks for each Data Point

Display tick marks for each data point along the horizontal axis.

Bars and Sunflower Plot Tab

Bars from Data Points to Horizontal (Vertical) Axis

A line or bar can be displayed that will connect each point to the horizontal or vertical axes. This option is useful if you want to generate a “bar chart” look to your scatter plot. Such a plot emphasizes the horizontal and vertical differences among the data points.



Bar Direction

This option controls where the bar extend to.

- **None**

  No bars are drawn.

- **Down**

  Horizontal bars are drawn from the points down to the bottom of the plot.

- **Up**

  Horizontal bars are drawn up from the data points to the top of the plot.

- **Left**
  Vertical bars are drawn from the points to the left axis.

- **Right**
  Vertical bars are drawn from the data points to the right axis.

- **Both**
  Bars are drawn in both directions.

Bars from Data Points to Horizontal (Vertical) Axis – Fill

Fill Color and Pattern

These options control the fill colors and patterns of the inside of the bars.

Bars from Data Points to Horizontal (Vertical) Axis – Outline

Outline Color and Width

These options control the color and width of the outline of the bar.

Bars from Data Points to Horizontal (Vertical) Axis – Width

Select Bar Width Parameter

Specify the method used to set the width of the bars as either Amount or Percent Space.

Amount

Designates a specific width for each bar.

Percent Empty Space

The percent of the horizontal axis length that is space instead of bars. This value determines the width of the bars. The smaller this value, the larger the bar width. Also note that this parameter only works for non-overlapping bars.

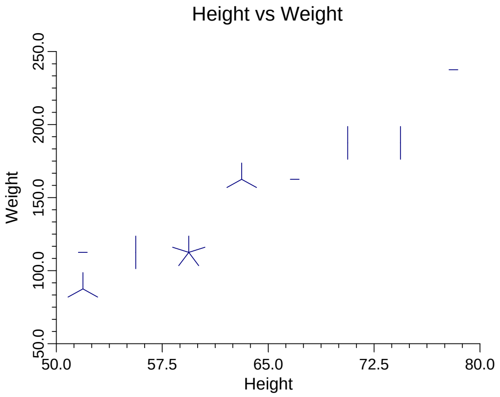
Sunflower Plot

The sunflower plot is used when you have so many observations that all you see on your scatter plot is a blob. Because of the large amount of overprinting, you cannot see many of the subtle patterns that occur in your data. Also, the eye tends to concentrate on outliers rather than the body of the data. The sunflower plot summaries the scatter plot by grouping the data so that you only see a few plot points. The algorithm is as follows:

1. Partition the plot with a two-way grid. The number of rows and columns in the grid determines the degree of smoothing.

2. Count the number of points in each cell of the grid. This is the amount that is plotted.

3. Plot a flower according to the following rules: If there is only one point in the cell, plot a point in the center of the cell. If there are more than one observation, make a "flower" with one petal for each point. Arrange the petals evenly about the center of the cell.



Count Pattern

This option specifies whether the sunflowers are displayed and which type of plot to display. Note that the data points will be displayed on top of the sunflowers unless you omit them by changing the Symbol 1 to None.

- **None**

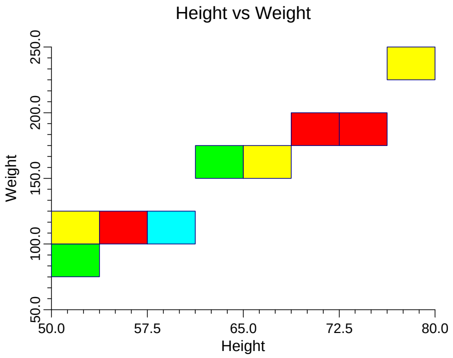
  Selecting this option omits the sunflowers from the scatter plot.

- **Spokes**

  Selecting this option requests the display of the standard sunflower plot.

- **Background Colors**

  Selecting this option requests a variation of the sunflower plot in which the sunflowers are not shown. Instead, the cells of the grid are displayed as rectangles in various colors. The cell's color is determined by the number of points that fall into it.



Vertical Bins

The number of vertical bins (slices or intervals) that are used. You will have to try a few different values for each plot to find the one that best represents the data.

Horizontal Bins

The number of horizontal bins (slices or intervals) that are used.

Petal Length

This is a percentage adjustment to the length of the petal. A value of 100 here indicates that the petals are to run to the edge of the cell. Percentage values of less than 100 reduce the length of the petals. We have found that a value of 90 works well in most cases.

Maximum Petals

Your data may contain cells that contain hundreds of data points. This option lets you specify a maximum number of petals. Cells with a count greater than this number still display only this many petals.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Scatter Plot

This section presents an example of how to generate a simple scatter plot. The data used are from the SAMPLE database. We will create a scatter plot of variables *Weight* versus *Height*.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Scatter Plots window.

1 Open the SAMPLE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.

2 Open the Scatter Plots window.

- On the menus, select **Graphics**, then **Scatter Plots**. The Scatter Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

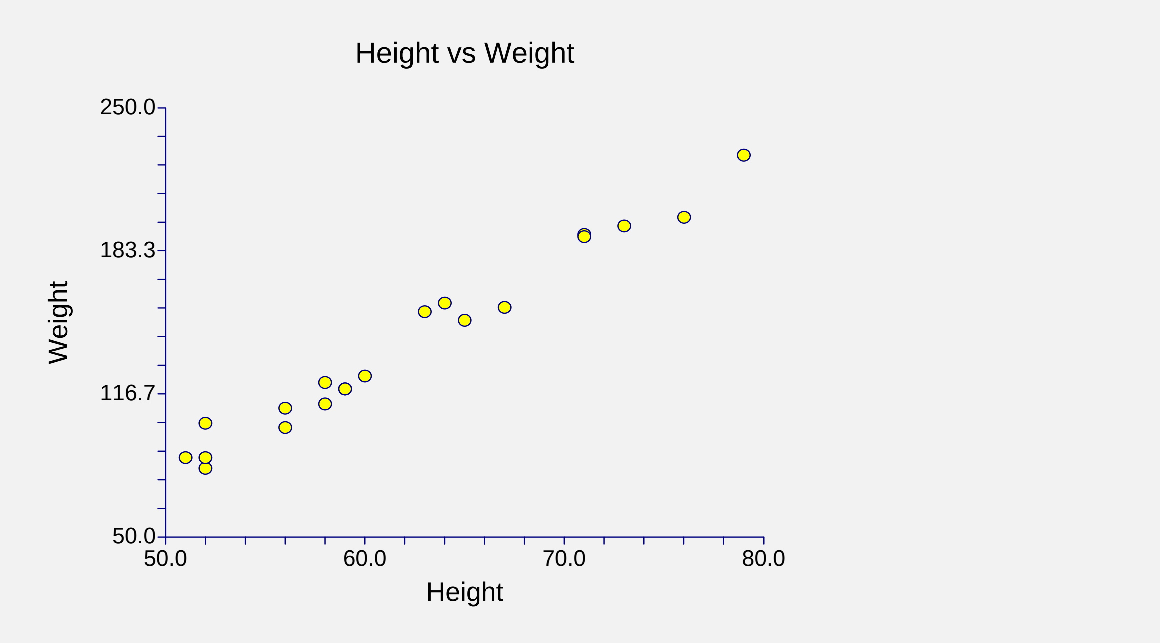
3 Specify the variables.

- On the Scatter Plots window, select the **Variables tab**.
- Double-click in the **Vertical Variable(s)** text box. This will bring up the variable selection window.
- Select **Weight** from the list of variables and then click **Ok**. “Weight” will appear in the Vertical Variable(s) box.
- Double-click in the **Horizontal Variable(s)** text box. This will bring up the variable selection window.
- Select **Height** from the list of variables and then click **Ok**. “Height” will appear in the Horizontal Variable(s) box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Scatter Plot Output



Creating a Scatter Plot Style File

One of the most exciting features of **NCSS** is its ability to mix graphics and text on the output reports. Many of the statistical procedures include scatter plots, box plots, or other graphs as part of their reports. Each of these graphs have over 200 options, so adding the options necessary to specify each graph would greatly increase the number of options that you would have to specify.

To overcome this, we let you create and save scatter plot style files. These style files contain the current settings of all options. When you use the style file in another procedure, such as the Multiple Regression, you only have to set a few of the options. Most of the options come from this style file. A default scatter plot style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a Scatter Plot Style file.

1 Open the SAMPLE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.
- Note: You do not necessarily have to use the SAMPLE database. You can use whatever database is easiest for you. Just open the appropriate database here.

2 Open the Scatter Plots window.

- On the menus, select **Graphics**, then **Scatter Plots**. The Scatter Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the horizontal and vertical variables.

- On the Scatter Plots window, select the **Variables tab**.
- Double-click in the **Vertical Variable(s)** text box. This will bring up the variable selection window.
- Select **Weight** from the list of variables and then click **Ok**. “Weight” will appear in the Vertical Variable(s) box.
- Double-click in the **Horizontal Variable(s)** text box. This will bring up the variable selection window.
- Select **Height** from the list of variables and then click **Ok**. “Height” will appear in the Horizontal Variable(s) box.
- Note: don’t worry about the specification of these variables. The actual variable names are not stored in the style file. They are used here so that you can see what your style will look like.

4 Set your options.

- Set the various options of the scatter plot’s appearance to the way you want them.
- Run the procedure to generate the scatter plot. This gives you a final check on whether the scatter plot appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 **Save the template (optional).**

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template is not the style file.
- To store the template, select the **Template tab** on the Scatter Plots window.
- Enter an appropriate name in the **File Name** box.
- Enter an appropriate phrase at the bottom of the window in the **Template Id** (the long box across the bottom of the Scatter Plot's window). This phrase will be displayed in the Template Id's box to help you identify the template files.
- Select **Save Template** from the File menu. This will save the template.

6 **Create and Save the Style File.**

- Select **Save Style File** from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the **Selected File** box. You can either reuse one of the style files that already exist or create a new name. You don't have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 **Using a Style File.**

- Using the style file is easy. For example, suppose you want to use this style file to plot residuals in the Multiple Regression procedure. You do the following:
- Select the **Plot Options tab** in the Multiple Regression procedure.
- Click the button to the right of the **Probability Plot - Plot Style File** box (the initial file name is Default). This will bring up the Scatter Plot Style File Selection window.
- Click on the appropriate file so that it is listed in the Selected File box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Multiple Regression window. That's it. Your new style has been activated.

Chapter 180

Color Selection Window

Introduction

The Color Selection window lets you choose an appropriate color from the 16 million colors that are available on today's monitors using the RGB color space. Although choosing a color sounds like a trivial task, it can become time-consuming and frustrating. When you have invested a lot of time and money in a project and now have important results to communicate, you probably want to take the time to make outstanding graphics. A few, well-chosen charts can communicate results quickly and effectively. An important feature of a chart is the color scheme that you use. The goal of the color selection window is to provide a tool that will allow you to pick a set of colors that are pleasing to the eye when viewed together, and let the viewer interpret the results quickly and effectively.

Color Theory

Picking colors is much easier once you discover the simple mathematical rules that should be followed. These rules are based on the theory of color. We suggest you browse the online article on color theory provided at **www.worqx.com**. This site provides a great introduction to color theory and color selection.

Color Models

In the online article, various color models are discussed. Computer users are familiar with the RGB (red, green, blue) model, which gives all the colors that are available on a computer monitor. Printers are more familiar with the CYM or CYMK model. Many other models have been developed. Our favorite model for choosing colors is the HSB (hue, saturation, brightness) model. When using this model, first pick the basic color or hue. Next, select the saturation and brightness of the color. Once a color has been selected, matching colors can be found by selecting other hues while keeping the saturation and brightness the same.

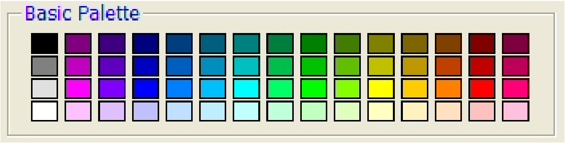
Color Wheel

The task of choosing several matching colors is aided by the use of a color wheel. A color wheel shows several hues all with the same saturation and brightness. Hence, all of the colors on the color wheel 'match'. The colors are arranged on the wheel so that colors on opposite sides of the wheel have high contrast, while adjacent colors have low contrast. Usually, high-contrasting colors are desirable when representing different treatment levels.

The Color Selection Window

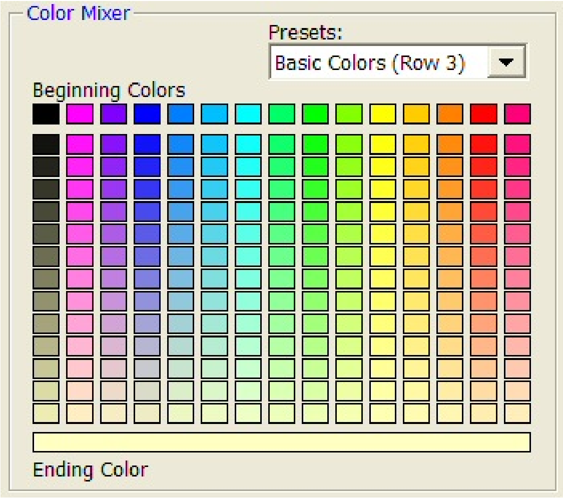
The Color Selection Window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette



These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



Presets: Basic Colors (Row 3)

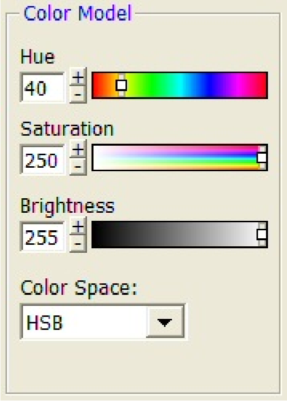
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model



Hue 40

Saturation 250

Brightness 255

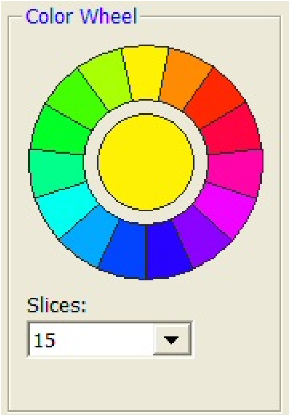
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



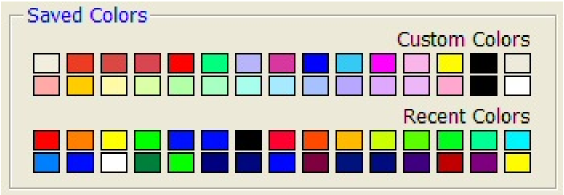
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors

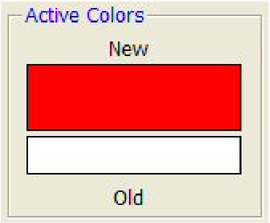


The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Active Colors



This box displays the original (Old) color and the selected (New) color.

The resulting color can be set to the original color by dragging and dropping the Old color on the New color.

Either of these colors can be dragged/dropped to any other color box.

180-4 Color Selection Window

Chapter 181

Symbol Settings Window

Introduction

The Symbol Settings window specifies the characteristics of a plotting symbol. The options are placed under two tabs. The first (Color) tab contains the options for specifying the colors of the symbol. The second (Symbol Settings) tab contains the options for specifying the type and size of the symbol.

Window Options

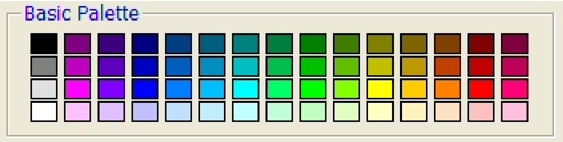
Color Tab

The Color tab lets you pick the fill and border colors of the plot symbol. The window is made up of several components that allow you to pick colors. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Clicks Modify Fill / Border

This option indicates whether color selections apply to the interior (fill) or border of the symbol.

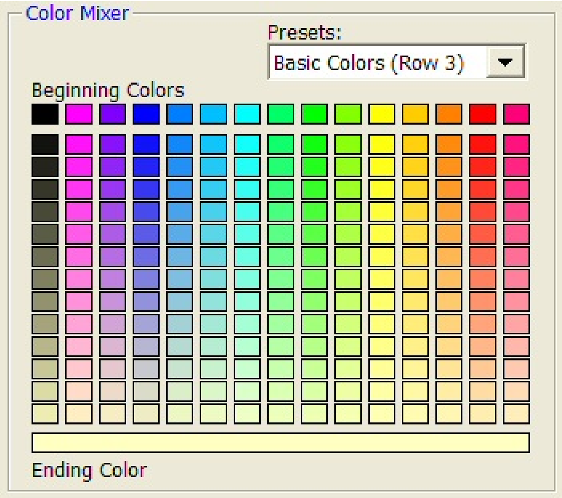
Basic Palette



These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

181-2 Symbol Settings Window

Color Mixer



Presets: Basic Colors (Row 3)

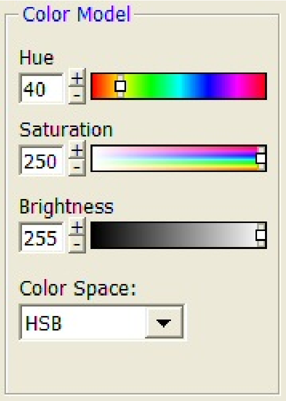
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model



Hue 40

Saturation 250

Brightness 255

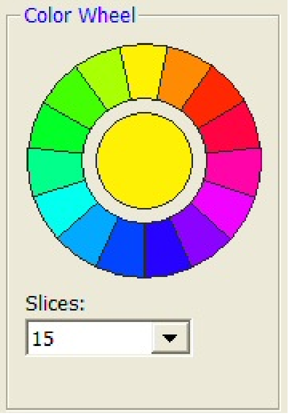
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



Slices: 15

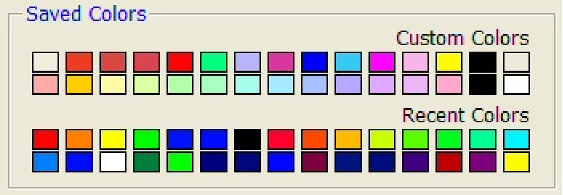
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Symbol Settings Tab

The Symbol Settings tab selects various attributes of the symbol.

Symbol Appearance

Type

This option specifies the type of symbol. If the symbol type only uses one color (such as a letter), the border color is used.

Radius

The radius specifies the size of the symbol. The default value of 100 works well in most cases.

Fill

Pattern

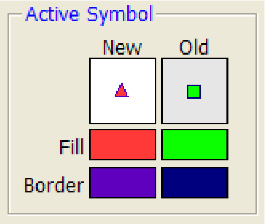
This is the pattern used to fill the interior of a solid symbol. We recommend setting this option to 'solid'.

Border

Width

This option specifies the width of the border.

Active Symbol



This box displays the symbol using the original (Old) settings and the selected (New) settings. Any of the colors can be dragged/dropped to any other color box.

Chapter 182

Text Settings Window

Introduction

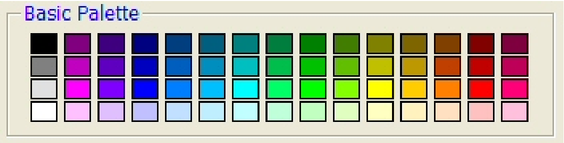
The Text Settings window specifies the characteristics (color and format) of a line of text, such as a label or title. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Text Settings) tab contains the options for specifying the format of the text.

Window Options

Color Tab

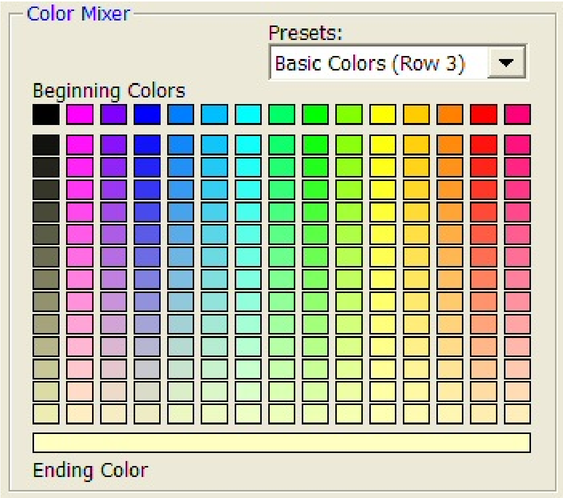
The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette



These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



Presets: Basic Colors (Row 3)

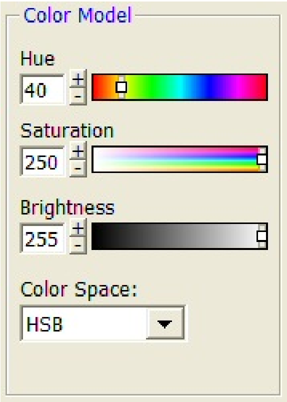
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model



Hue: 40

Saturation: 250

Brightness: 255

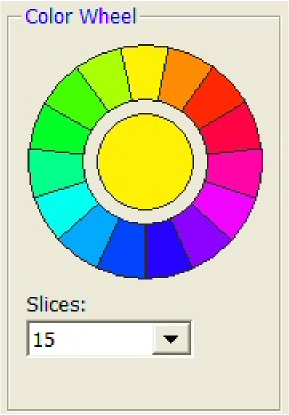
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



Slices: 15

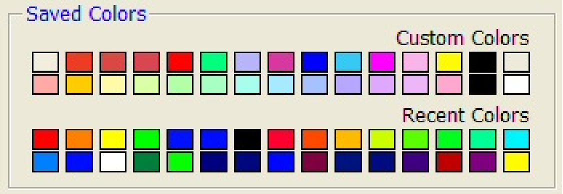
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Text Settings Tab

The Text Settings tab selects various attributes of the title or label.

Text Attributes

Font Size

This option sets the font size of the text. The minimum font size is '1'.

Bold, Italic, and Underline

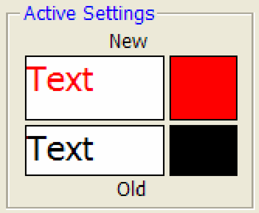
Checking any of these items causes the text to have the corresponding attribute.

Text Value

Actual Text

This option sets the actual text that is to be displayed.

Active Settings



This box displays the original (Old) text format and the selected (New) text format.

The resulting format can be set to the original format by dragging and dropping the Old format on the New format.

Either of these colors can be dragged/dropped to any other color box

182-4 Text Settings Window

Chapter 183

Line Settings Window

Introduction

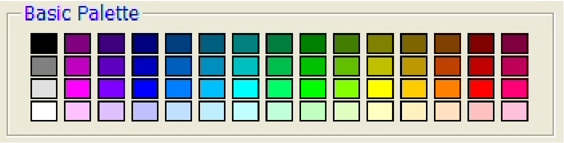
The Line Settings window specifies the characteristics (color, width, and pattern) of a line. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Settings) tab contains the options for specifying the line width and pattern.

Window Options

Color Tab

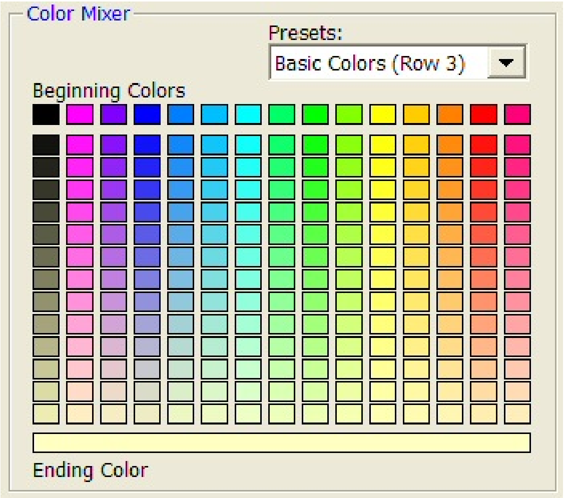
The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette



These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



Presets: Basic Colors (Row 3)

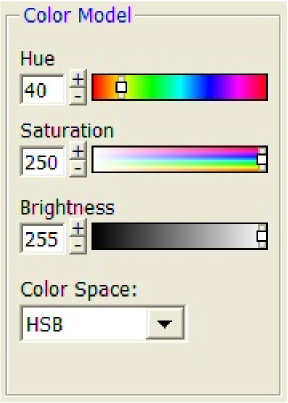
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model



Hue 40

Saturation 250

Brightness 255

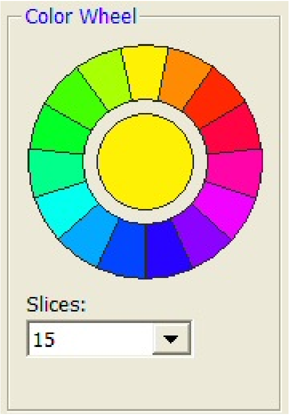
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



Slices: 15

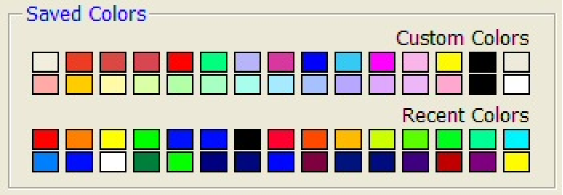
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Settings Tab

The Settings tab selects the width and pattern of the line.

Specify Line Settings

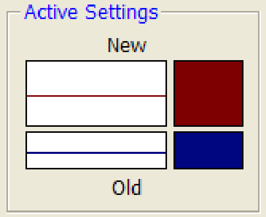
Width

This option sets the width of the line. The minimum line width is '1'.

Line Pattern

This option sets the pattern of the line. Patterns other than 'solid' require the width to be less than 30. Note that the line pattern was more useful in the days before color printers became common. Nowadays, non-solid lines are used less frequently.

Active Settings



This box displays the original (Old) line color and the selected (New) line color.

The resulting line can be set to the original line by dragging and dropping the Old line on the New line.

Either of these colors can be dragged/dropped to any other color box.

183-4 Line Settings Window

Chapter 184

Axis-Line Settings Window

Introduction

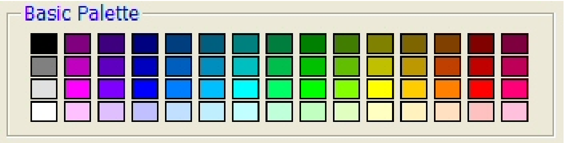
The Axis-Line Settings window specifies the characteristics (color and width) of the line used as the axis. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Width & Type) tab contains the options for specifying the line width and type.

Window Options

Color Tab

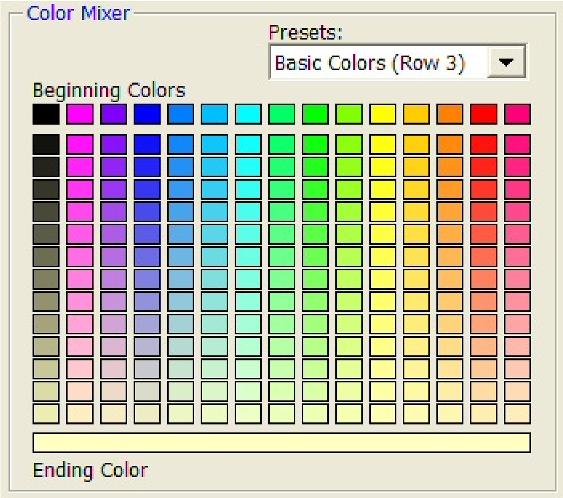
The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette



These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



Presets: Basic Colors (Row 3)

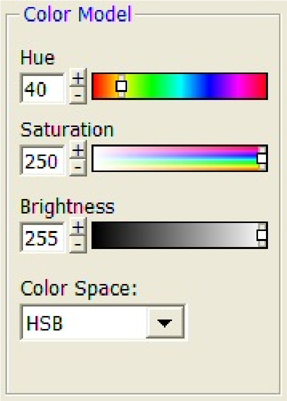
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model



Hue: 40

Saturation: 250

Brightness: 255

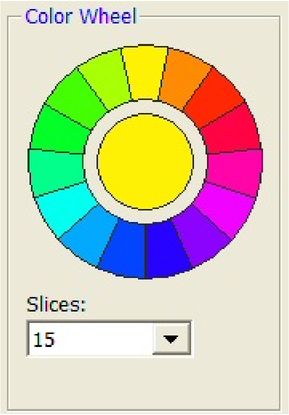
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



Slices: 15

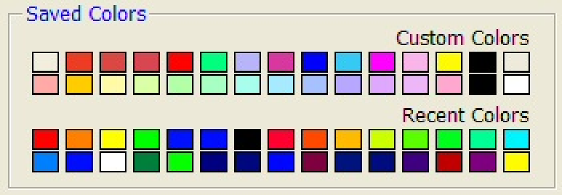
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Width & Type Tab

The Width & Type tab selects the width and pattern of an axis line.

Specify Width

Width

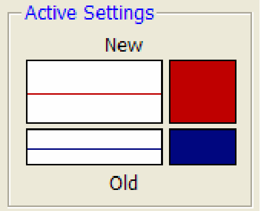
This option sets the width of the line. The minimum line width is '1'.

Specify Type

Type

This option whether the axis is text or numeric. When Text is selected, each unique value is equally-spaced along the axis.

Active Settings



This box displays the original (Old) line color and the selected (New) line color.

The resulting line can be set to the original line by dragging and dropping the Old line on the New line.

Either of these colors can be dragged/dropped to any other color box.

Chapter 185

Grid / Tick Settings Window

Introduction

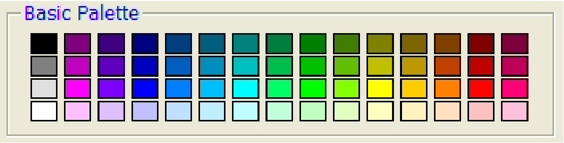
The Grid / Tick Settings window specifies the characteristics (color and format) of the tickmarks displayed along the axis as well as the grid lines in the body of the plot. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Grid & Tick Settings) tab contains the options for specifying the format of the grid lines and tickmarks.

Window Options

Color Tab

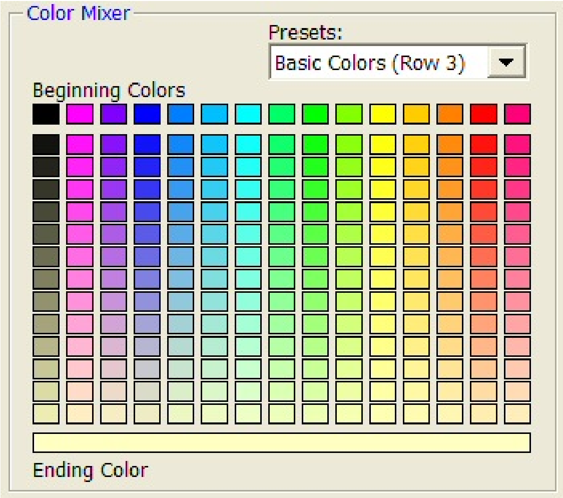
The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette



These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer


Color Mixer

Presets: Basic Colors (Row 3)

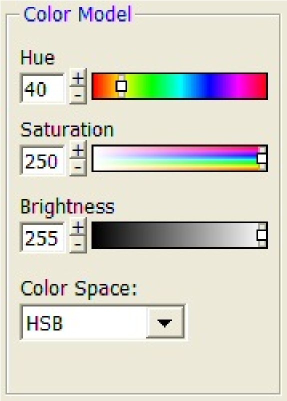
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model


Color Model

Hue: 40

Saturation: 250

Brightness: 255

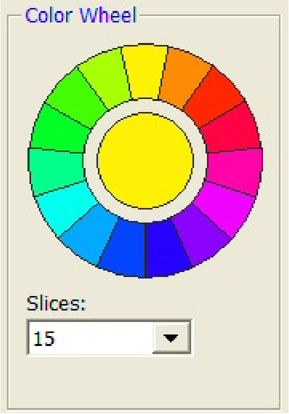
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel


Color Wheel

Slices: 15

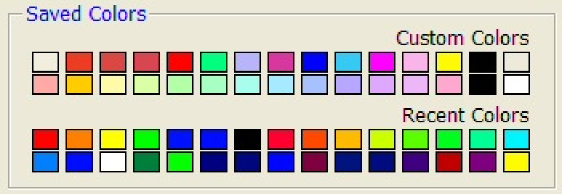
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Grid & Tick Settings Tab

The Grid & Tick Settings tab sets various attributes of the grid lines and tickmarks.

Grid Settings

Grid Width

This option sets the width of the grid lines. The minimum width is '1'.

Grid Pattern

This option sets the pattern of the grid line. Patterns other than 'solid' require the width to be less than 30. Note that the grid line pattern was more useful in the days before color printers became common. Nowadays, non-solid lines are used less frequently.

Tick Settings

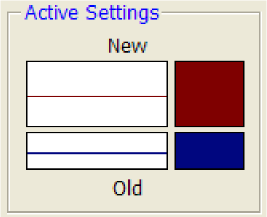
Tick Width

This option sets the width (size) of the tickmarks. The minimum width is '1'.

Tick Length

This option determines the length of the tickmarks.

Active Settings



This box displays the original (Old) grid line and tick color and the selected (New) grid line and tick color.

The resulting line and color can be set to the original line and color by dragging and dropping the Old line onto the New line.

Either of these colors can be dragged/dropped to any other color box.

185-4 Grid / Tick Settings Window

Chapter 186

Tick Label Settings Window

Introduction

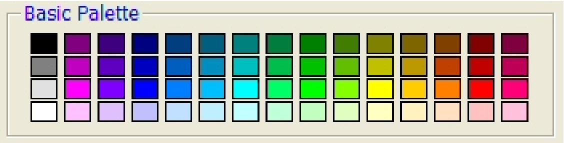
The tick label settings window specifies the characteristics (color and format) of the reference numbers displayed at the tickmarks along the axis. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Text Settings) tab contains the options for specifying the format of the reference numbers.

Window Options

Color Tab

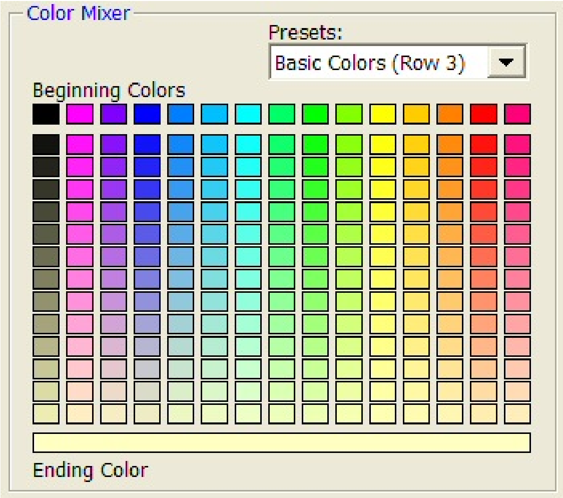
The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette



These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



Presets: Basic Colors (Row 3)

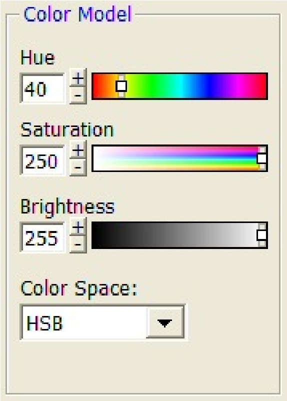
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model



Hue: 40

Saturation: 250

Brightness: 255

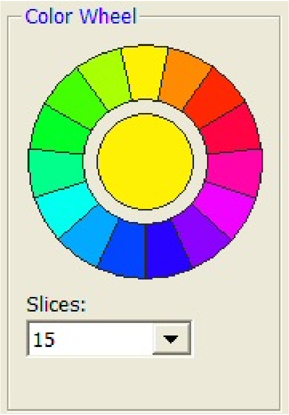
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



Slices: 15

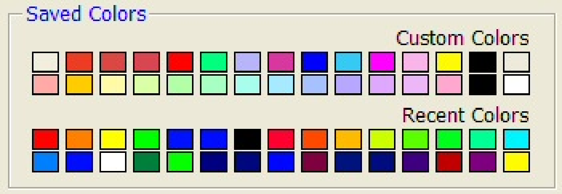
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Text Settings Tab

The Text Settings tab selects various attributes of the reference numbers.

Text Attributes

Font Size

This option sets the font size of the number. The minimum font size is '1'.

Bold, Italic, and Underline

Checking any of these items causes the text to have the corresponding attribute.

Text Format

Decimals

This option determines the number of decimal places that are displayed.

Max Characters

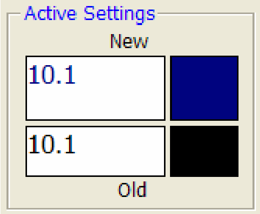
This option determines how much space is provided for displaying the reference numbers.

Text Rotation

Horizontal and Vertical

This option determines whether the text is displayed horizontally or vertically.

Active Settings



This box displays the original (Old) reference number format and the selected (New) reference number format.

The resulting format can be set to the original format by dragging and dropping the Old format on the New format.

Either of these colors can be dragged/dropped to any other color box.

186-4 Tick Label Settings Window

Chapter 187

Heat Map Settings Window

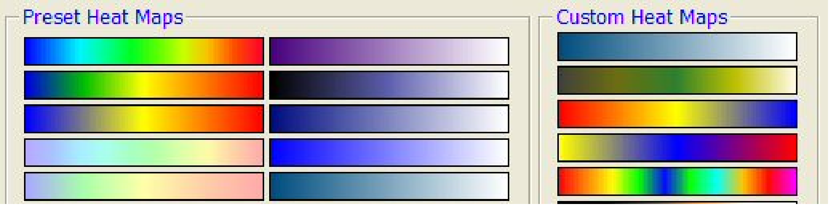
Introduction

The Heat Map Settings window specifies the characteristics of a heat map. The options are placed under three tabs. The first (Heat Maps) tab contains several preconfigured heat maps. The second (Color Selection) tab contains the options for specifying the colors of the heat map. The third (Intervals & Scaling) tab contains the options for specifying the number of color-intervals and the scaling of the heat map.

Window Options

Heat Maps Tab

The Heat Maps tab displays several preset and heat maps heat map patterns that can be selected.



Preset Heat Maps

We have provided several heat map patterns, which show the rich heat map variety that is possible. Simply click on a heat map to activate it. You will see it appear as the selected heat map in the upper right of the window.

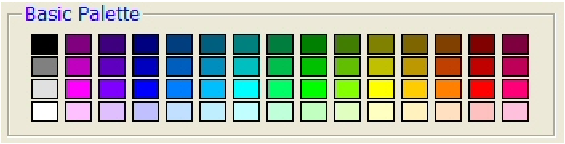
Custom Heat Maps

The custom heat maps are heat maps that you create that you want to save for later use. They are stored and will be available until you change them.

Color Selection Tab

The Color Selection tab lets you pick the colors to be blended to form a heat map. The window is made up of several components that allow you to pick colors. We will now describe each of these components in turn.

Basic Palette

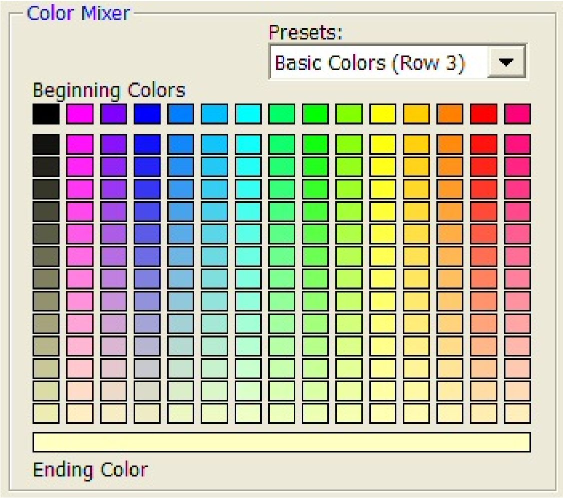


These boxes hold a set of common colors. To use a color, click it and change the currently active color in the Heat Map colors frame or drag/drop it to one of the Heat Map colors.

Active Color

This box provides a large view of the currently-selected color. As you select different colors, this window will change.

Color Mixer



Presets: Basic Colors (Row 3)

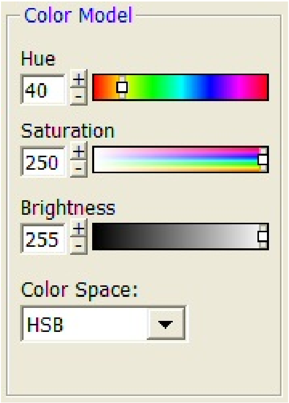
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model



Hue 40

Saturation 250

Brightness 255

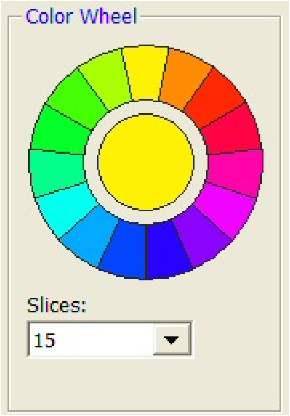
Color Space: HSB

This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



Slices: 15

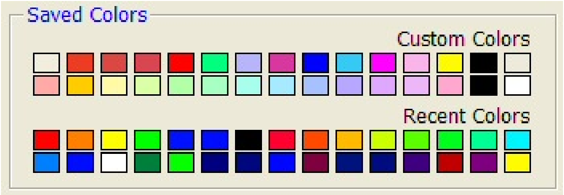
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Intervals & Scaling Tab

The Intervals & Scaling tab has options that determine the intervals and scale of the heatmap.

Number of Intervals

Intervals

This option specifies the number of intervals along the heat map. The default value of 100 seems to work well in most cases.

Scale Along the Heat Map Axis

Scale

This option specifies the type of scaling that is used.

- **Regular**

  The range from the minimum to the maximum is divided up into equal-width intervals.

- **Percentile**

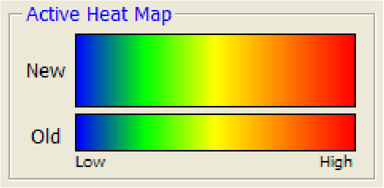
  The range from the minimum to the maximum is divided into percentiles. If the number of intervals is set to 100, then the values in the first percentile receive the first color, the values in the second percentile receive the second color, and so on.

  This scale method forces all colors of the heat map to be displayed. Care must be used when interpreting the data since the intervals are not equally spaced.

- **Log**

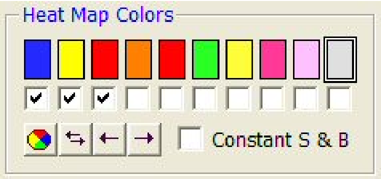
  The values are displayed on according to a logarithmic scale.

Active Heat Map



This box displays the heat map using the original (Old) settings and the selected (New) settings. This heat map can be dragged/dropped to any of the Custom Heat Maps under the Heat Maps tab.

Heat Map Colors



This frame displays the colors that are used to form the heat map. The gray box on the right has a border around it, which indicates that if a color box is clicked, it will change the color of this box.

Even though a color appears in one of these boxes, it will not be used in the heat map unless the check box immediate beneath it is checked.

A row of special buttons appears below the check boxes. We will now describe each of these buttons.

Color Wheel Button

Clicking the first button causes the colors displayed on the Color Wheel to be transferred to these boxes.

Reverse Order Button

Clicking the second button causes the order of the buttons to be reversed.

Shift Left Button

Clicking the third button causes the colors to be shifted one box to the left.

Shift Right Button

Clicking the fourth button causes the colors to be shifted one box to the right.

Constant S & B

Checking this check box causes all of the boxes to by reset so that they have the same saturation and brightness as the highlighted box, while maintaining the same hue.

187-6 Heat Map Settings Window

References

A

Agresti, A. and Coull, B. 1998. “Approximate is Better than ‘Exact’ for Interval Estimation of Binomial Proportions,” *American Statistician*, Volume 52 Number 2, pages 119-126.

A’Hern, R. P. A. 2001. "Sample size tables for exact single-stage phase II designs." *Statistics in Medicine*, Volume 20, pages 859-866.

AIAG (Automotive Industry Action Group). 1995. *Measurement Systems Analysis*. This booklet was developed by Chrysler/Ford/GM Supplier Quality Requirements Task Force. It gives a detailed discussion of how to design and analyze an R&R study. The book may be obtained from ASQC or directly from AIAG by calling 801-358-3570.

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle," In B. N. Petrov & F. Csaki (Eds.), *The second international symposium on information theory*. Budapest, Hungary: Akademiai Kiado.

Akaike, H. 1974. "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, (6): pages 716-723.

Albert, A. and Harris, E. 1987. *Multivariate Interpretation of Clinical Laboratory Data*. Marcel Dekker, New York, New York. This book is devoted to a discussion of how to apply multinomial logistic regression to medical diagnosis. It contains the algorithm that is the basis of our multinomial logistic regression routine.

Allen, D. and Cady, F.. 1982. *Analyzing Experimental Data by Regression*. Wadsworth. Belmont, Calif. This book works completely through several examples. It is very useful to those who want to see complete analyses of complex data.

Al-Sunduqchi, Mahdi S. 1990. *Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics*. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.

Altman, Douglas. 1991. *Practical Statistics for Medical Research*. Chapman & Hall. New York, NY. This book provides an introductory discussion of many statistical techniques that are used in medical research. It is the only book we found that discussed ROC curves.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. 1997. *Statistical Models Based on Counting Processess*. Springer-Verlag, New York. This is an advanced book giving many of the theoretically developments of survival analysis.

Anderson, R.L. and Hauck, W.W. 1983. "A new Procedure for testing equivalence in comparative bioavailability and other clinical trials." *Commun. Stat. Theory Methods.*, Volume 12, pages 2663-2692.

Anderson, T.W. and Darling, D.A. 1954. "A test of goodness-of-fit." *J. Amer. Statist. Assoc*, Volume 49, pages 765-769.

Andrews, D.F., and Herzberg, A.M. 1985. *Data*. Springer-Verlag, New York. This book is a collection of many different data sets. It gives a complete description of each.

Armitage. 1955. “Tests for linear trends in proportions and frequencies.” *Biometrics*, Volume 11, pages 375-386.

Armitage, P., and Colton, T. 1998. *Encyclopedia of Biostatistics*. John Wiley, New York.

Armitage,P., McPherson, C.K., and Rowe, B.C. 1969. “Repeated significance tests on accumulating data.” *Journal of the Royal Statistical Society, Series A*, 132, pages 235-244.

Atkinson, A.C. 1985. *Plots, Transformations, and Regression*. Oxford University Press, Oxford (also in New York). This book goes into the details of regression diagnostics and plotting. It puts together much of the recent work in this area.

Atkinson, A.C., and Donev, A.N. 1992. *Optimum Experimental Designs*. Oxford University Press, Oxford. This book discusses D-Optimal designs.

Austin, P.C., Grootendorst, P., and Anderson, G.M. 2007. “A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study,” *Statistics in Medicine*, Volume 26, pages 734-753.

B

Bain, L.J. and Engelhardt, M. 1991. *Statistical Analysis of Reliability and Life-Testing Models*. Marcel Dekker. New York. This book contains details for testing data that follow the exponential and Weibull distributions.

Baker, Frank. 1992. *Item Response Theory*. Marcel Dekker. New York. This book contains a current overview of IRT. It goes through the details, providing both formulas and computer code. It is not light reading, but it will provide you with much of what you need if you are attempting to use this technique.

Barnard, G.A. 1947. "Significance tests for 2 x 2 tables." *Biometrika* 34:123-138.

Barrentine, Larry B. 1991. *Concepts for R&R Studies*. ASQC Press. Milwaukee, Wisconsin. This is a very good applied work book on the subject of repeatability and reproducibility studies. The ISBN is 0-87389-108-2. ASQC Press may be contacted at 800-248-1946.

Bartholomew, D.J. 1963. "The Sampling Distribution of an Estimate Arising in Life Testing." *Technometrics*, Volume 5 No. 3, 361-374.

Bartlett, M.S. 1950. "Tests of significance in factor analysis." *British Journal of Psychology (Statistical Section)*, 3, 77-85.

Bates, D. M. and Watts, D. G. 1981. "A relative offset orthogonality convergence criterion for nonlinear least squares," *Technometrics*, Volume 23, 179-183.

Beal, S. L. 1987. "Asymptotic Confidence Intervals for the Difference between Two Binomial Parameters for Use with Small Samples." *Biometrics*, Volume 43, Issue 4, 941-950.

Belsley, Kuh, and Welsch. 1980. *Regression Diagnostics*. John Wiley & Sons. New York. This is the book that brought regression diagnostics into the main-stream of statistics. It is a graduate level treatise on the subject.

Benjamini, Y. and Hochberg, Y. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society, Series B (Methodological),* Vol. 57, No. 1, 289-300.

Bertsekas, D.P. 1991. *Linear Network Optimization: Algorithms and Codes*. MIT Press. Cambridge, MA.

Blackwelder, W.C. 1993. "Sample size and power in prospective analysis of relative risk." *Statistics in Medicine*, Volume 12, 691-698.

Blackwelder, W.C. 1998. "Equivalence Trials." In *Encyclopedia of Biostatistics*, John Wiley and Sons. New York. Volume 2, 1367-1372.

Bloomfield, P. 1976. *Fourier Analysis of Time Series*. John Wiley and Sons. New York. This provides a technical introduction to fourier analysis techniques.

Bock, R.D., Aiken, M. 1981. "Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, 46, 443-459.

Bolstad, B.M., et al. 2003. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 19, 185-193.

Bonett, Douglas. 2002. “Sample Size Requirements for Testing and Estimating Coefficient Alpha.” *Journal of Educational and Behavioral Statistics*, Vol. 27, pages 335-340.

Box, G.E.P. and Jenkins, G.M. 1976. *Time Series Analysis - Forecasting and Control*. Holden-Day.: San Francisco, California. This is the landmark book on ARIMA time series analysis. Most of the material in chapters 6 - 9 of this manual comes from this work.

Box, G.E.P. 1949. "A general distribution theory for a class of likelihood criteria." *Biometrika,* 1949, **36**, 317-346.

Box, G.E.P. 1954a. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: I." *Annals of Mathematical Statistics*, **25**, 290-302.

Box, G.E.P. 1954b. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: II." *Annals of Mathematical Statistics*, **25**, 484-498.

Box, G.E.P., Hunter, S. and Hunter. 1978. *Statistics for Experimenters*. John Wiley & Sons, New York. This is probably the leading book in the area experimental design in industrial experiments. You definitely should acquire and study this book if you plan anything but a casual acquaintance with experimental design. The book is loaded with examples and explanations.

Breslow, N. E. and **Day, N. E.** 1980. *Statistical Methods in Cancer Research: Volume 1. The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

Brown, H., and Prescott, R. 2006. *Applied Mixed Models in Medicine*. 2nd ed. John Wiley & Sons Ltd. Chichester, West Sussex, England.

Brush, Gary G. 1988. *Volume 12: How to Choose the Proper Sample Size*, American Society for Quality Control, 310 West Wisconsin Ave, Milwaukee, Wisconsin, 53203. This is a small workbook for quality control workers.

Burdick, R.K. and Larsen, G.A. 1997. “Confidence Intervals on Measures of Variability in R&R Studies.” *Journal of Quality Technology, Vol. 29, No. 3, Pages 261-273*. This article presents the formulas used to construct confidence intervals in an R&R study.

Bury, Karl. 1999. *Statistical Distributions in Engineering.*. Cambridge University Press. New York, NY. (www.cup.org).

C

Cameron, A.C. and Trivedi, P.K. 1998. *Regression Analysis of Count Data*. Cambridge University Press. New York, NY. (www.cup.org).

Carmines, E.G. and Zeller, R.A. 1990. *Reliability and Validity Assessment*. Sage University Paper. 07-017. Newbury Park, CA.

Casagrande, J. T., Pike, M.C., and Smith, P. G. 1978. “The Power Function of the “Exact” Test for Comparing Two Binomial Distributions,” *Applied Statistics*, Volume 27, No. 2, pages 176-180. This article presents the algorithm upon which our Fisher’s exact test is based.

Cattell, R.B. 1966. "The scree test for the number of factors." *Mult. Behav. Res.* 1, 245-276.

Cattell, R.B. and Jaspers, J. 1967. "A general plasmode (No. 30-10-5-2) for factor analytic exercises and research." *Mult. Behav. Res. Monographs*. 67-3, 1-212.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. 1983. *Graphicals Methods for Data Analysis.* Duxbury Press, Boston, Mass. This wonderful little book is full of examples of ways to analyze data graphically. It gives complete (and readable) coverage to such topics as scatter plots, probability plots, and box plots. It is strongly recommended.

Chatfield, C. 1984. *The Analysis of Time Series*. Chapman and Hall. New York. This book gives a very readable account of both ARMA modeling and spectral analysis. We recommend it to those who wish to get to the bottom of these methods.

Chatterjee and Price. 1979. *Regression Analysis by Example*. John Wiley & Sons. New York. A great hands-on book for those who learn best from examples. A newer edition is now available.

Chen, K.W.; Chow, S.C.; and Li, G. 1997. "A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs" *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 25, No. 6, pages 753-765.

Chen, T. T. 1997. "Optimal Three-Stage Designs for Phase II Cancer Clinical Trials." *Statistics in Medicine*, Volume 16, pages 2701-2711.

Chen, Xun. 2002. "A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases." *Statistics in Medicine*, Volume 21, pages 943-956.

Chow, S.C. and Liu, J.P. 1999. *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker. New York.

Chow, S.C.; Shao, J.; Wang, H. 2003. *Sample Size Calculations in Clinical Research*. Marcel Dekker. New York.

Cochran and Cox. 1992. *Experimental Designs. Second Edition*. John Wiley & Sons. New York. This is one of the classic books on experimental design, first published in 1957.

Cochran, W.G. and Rubin, D.B. 1973. “Controlling bias in observational studies,” *Sankhya, Ser. A*, Volume 35, Pages 417-446.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, New Jersey. This is a very nice, clearly written book. There are MANY examples. It is the largest of the sample size books. It does not deal with clinical trials.

Cohen, Jacob. 1990. “Things I Have Learned So Far.” *American Psychologist*, December, 1990, pages 1304-1312. This is must reading for anyone still skeptical about the need for power analysis.

Collett, D. 1991. *Modelling Binary Data.* Chapman & Hall, New York, New York. This book covers such topics as logistic regression, tests of proportions, matched case-control studies, and so on.

Collett, D. 1994. *Modelling Survival Data in Medical Research.* Chapman & Hall, New York, New York. This book covers such survival analysis topics as Cox regression and log rank tests.

Conlon, M. and Thomas, R. 1993. “The Power Function for Fisher’s Exact Test.” *Applied Statistics*, Volume 42, No. 1, pages 258-260. This article was used to validate the power calculations of Fisher’s Exact Test in PASS. Unfortunately, we could not use the algorithm to improve the speed because the algorithm requires equal sample sizes.

Conover, W.J. 1971. *Practical Nonparametric Statistics*. John Wiley & Sons, Inc. New York.

Conover, W.J., Johnson, M.E., and **Johnson, M.M.** 1981. *Technometrics***, 23,** 351-361**.**

Cook, D. and Weisberg, S. 1982. *Residuals and Influence in Regression*. Chapman and Hall. New York. This is an advanced text in the subject of regression diagnostics.

Cooley, W.W. and Lohnes, P.R. 1985. *Multivariate Data Analysis*. Robert F. Krieger Publishing Co. Malabar, Florida.

Cox, D. R. 1972. “Regression Models and life tables.” *Journal of the Royal Statistical Society, Series B,* Volume 34, Pages 187-220. This article presents the proportional hazards regression model.

Cox, D. R. 1975. “Contribution to discussion of Mardia (1975a).” *Journal of the Royal Statistical Society, Series B*, Volume 37, Pages 380-381.

Cox, D.R. and Snell, E.J. 1981. *Applied Statistics: Principles and Examples*. Chapman & Hall. London, England.

Cureton, E.E. and D'Agostino, R.B. 1983. *Factor Analysis - An Applied Approach*. Lawrence Erlbaum Associates. Hillsdale, New Jersey. (This is a wonderful book for those who want to learn the details of what factor analysis does. It has both the theoretical formulas and simple worked examples to make following along very easy.)

D

D'Agostino, R.B., Belanger, A., D'Agostino, R.B. Jr. 1990."A Suggestion for Using Powerful and Informative Tests of Normality.", *The American Statistician*, November 1990, Volume 44 Number 4, pages 316-321. This tutorial style article discusses D'Agostino's tests and tells how to interpret normal probability plots.

D'Agostino, R.B., Chase, W., Belanger, A. 1988."The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations.", *The American Statistician*, August 1988, Volume 42 Number 3, pages 198-202.

D'Agostino, R.B. Jr. 2004. *Tutorials in Biostatistics*. Volume 1. John Wiley & Sons. Chichester, England.

Dallal, G. 1986. "An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality," *The American Statistician*, Volume 40, Number 4, pages 294-296.

Daniel, C. and Wood, F. 1980. *Fitting Equations to Data*. John Wiley & Sons. New York. This book gives several in depth examples of analyzing regression problems by computer.

Daniel, W. 1990. *Applied Nonparametric Statistics.* 2nd ed. PWS-KENT Publishing Company. Boston.

Davies, Owen L. 1971. *The Design and Analysis of Industrial Experiments*. Hafner Publishing Company, New York. This was one of the first books on experimental design and analysis. It has many examples and is highly recommended.

Davis, J. C. 1985. *Statistics and Data Analysis in Geology*. John Wiley. New York. (A great layman's discussion of many statistical procedures, including factor analysis.)

Davison, A.C. and Hinkley, D.V. 1999. *Bootstrap Methods and their Applications*. Cambridge University Press. NY, NY. This book provides and detailed account of bootstrapping.

Davison, Mark. 1983. *Multidimensional Scaling*. John Wiley & Sons. NY, NY. This book provides a very good, although somewhat advanced, introduction to the subject.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. 1988. "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics*, 44, pages 837-845.

DeMets, D.L. and Lan, K.K.G. 1984. "An overview of sequential methods and their applications in clinical trials." *Communications in Statistics, Theory and Methods*, 13, pages 2315-2338.

DeMets, D.L. and Lan, K.K.G. 1994. "Interim analysis: The alpha spending function approach." *Statistics in Medicine*, 13, pages 1341-1352.

Demidenko, E. 2004. *Mixed Models – Theory and Applications*. John Wiley & Sons. Hoboken, New Jersey.

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. Academic Press. New York. (Presents many useful results for determining sample sizes.)

DeVor, Chang, and Sutherland. 1992. *Statistical Quality Design and Control*. Macmillan Publishing. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 800 pages.

Devroye, Luc. 1986. *Non-Uniform Random Variate Generation.* Springer-Verlag. New York. This book is currently available online at http://jeff.cs.mcgill.ca/~luc/rnbookindex.html.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. 1994. *Analysis of Longitudinal Data*. Oxford University Press. New York, New York.

Dillon, W. and Goldstein, M. 1984. *Multivariate Analysis - Methods and Applications*. John Wiley. NY, NY. This book devotes a complete chapter to loglinear models. It follows Fienberg's book, providing additional discussion and examples.

Dixon, W. J. and Tukey, J. W. 1968. "Approximate behavior of the distribution of Winsorized t," *Technometrics*, Volume 10, pages 83-98.

Dodson, B. 1994. *Weibull Analysis*. ASQC Quality Press. Milwaukee, Wisconsin. This paperback book provides the basics of Weibull fitting. It contains many of the formulas used in our Weibull procedure.

Donnelly, Thomas G. 1980. "ACM Algorithm 462: Bivariate Normal Distribution," *Collected Algorithms from ACM*, Volume II, New York, New York.

Donner, Allan. 1984. "Approaches to Sample Size Estimation in the Design of Clinical Trials--A Review," *Statistics in Medicine*, Volume 3, pages 199-214. This is a well done review of the clinical trial literature. Although it is becoming out of date, it is still a good place to start.

Donner, A. and Klar, N. 1996. "Statistical Considerations in the Design and Analysis of Community Intervention Trials." *The Journal of Clinical Epidemiology*, Vol. 49, No. 4, 1996, pages 435-439.

Donner, A. and Klar, N. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold. London.

Draghici, S. 2003. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC. London. This is an excellent overview of most areas of Microarray analysis.

Draper, N.R. and Smith, H. 1966. *Applied Regression Analysis*. John Wiley & Sons. New York. This is a classic text in regression analysis. It contains both in depth theory and applications. This text is often used in graduate courses in regression analysis.

Draper, N.R. and Smith, H. 1981. *Applied Regression Analysis - Second Edition*. John Wiley & Sons. New York, NY. This is a classic text in regression analysis. It contains both in-depth theory and applications. It is often used in graduate courses in regression analysis.

Dudoit, S., Shaffer, J.P., and Boldrick, J.C. 2003. "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, Volume 18, No. 1, pages 71-103.

Dudoit, S., Yang, Y.H., Callow, M.J., and **Speed, T.P.** 2002. "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Experiments," *Statistica Sinica*, Volume 12, pages 111-139.

du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H. 1986. *Graphical Exploratory Data Analysis.* Springer-Verlag. New York. This book contains examples of graphical analysis for a broad range of topics.

Dunn, O. J. 1964. "Multiple comparisons using rank sums," *Technometrics*, Volume 6, pages 241-252.

Dunnett, C. W. 1955. "A Multiple comparison procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, Volume 50, pages 1096-1121.

Dunteman, G.H. 1989. *Principal Components Analysis*. Sage University Papers, 07-069. Newbury Park, California. Telephone (805) 499-0721. This monograph costs only $7. It gives a very good introduction to PCA.

Dupont, William. 1988. "Power Calculations for Matched Case-Control Studies," *Biometrics*, Volume 44, pages 1157-1168.

Dupont, William and **Plummer, Walton D.** 1990. “Power and Sample Size Calculations--A Review and Computer Program,” *Controlled Clinical Trials*, Volume 11, pages 116-128. Documents a nice public-domain program on sample size and power analysis.

Durbin, J. and Watson, G. S. 1950. “Testing for Serial Correlation in Least Squares Regression - I,” *Biometrika*, Volume 37, pages 409-428.

Durbin, J. and Watson, G. S. 1951. “Testing for Serial Correlation in Least Squares Regression - II,” *Biometrika*, Volume 38, pages 159-177.

Dyke, G.V. and Patterson, H.D. 1952. “Analysis of factorial arrangements when the data are proportions.” *Biometrics*. Volume 8, pages 1-12. This is the source of the data used in the LLM tutorial.

E

Eckert, Joseph K. 1990. *Property Appraisal and Assessment Administration*. International Association of Assessing Officers. 1313 East 60th Street. Chicago, IL 60637-2892. Phone: (312) 947-2044. This is a how-to manual published by the IAAO that describes how to apply many statistical procedures to real estate appraisal and tax assessment. We strongly recommend it to those using our *Assessment Model* procedure.

Edgington, E. 1987. *Randomization Tests*. Marcel Dekker. New York. A comprehensive discussion of randomization tests with many examples.

Edwards, L.K. 1993. *Applied Analysis of Variance in the Behavior Sciences*. Marcel Dekker. New York. Chapter 8 of this book is used to validate the repeated measures module of PASS.

Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall. New York.

Elandt-Johnson, R.C. and Johnson, N.L. 1980. *Survival Models and Data Analysis*. John Wiley. NY, NY. This book devotes several chapters to population and clinical life-table analysis.

Epstein, Benjamin. 1960. “Statistical Life Test Acceptance Procedures.” *Technometrics*. Volume 2.4, pages 435-446.

Everitt, B.S. and Dunn, G. 1992. *Applied Multivariate Data Analysis*. Oxford University Press. New York. This book provides a very good introduction to several multivariate techniques. It helps you understand how to interpret the results.

F

Farrington, C. P. and Manning, G. 1990. “Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk.” *Statistics in Medicine*, Vol. 9, pages 1447-1454. This article contains the formulas used for the Equivalence of Proportions module in PASS.

Feldt, L.S.; Woodruff, D.J.; & Salih, F.A. 1987. “Statistical inference for coefficient alpha.” *Applied Psychological Measurement*, Vol. 11, pages 93-103.

Feldt, L.S.; Ankenmann, R.D. 1999. “Determining Sample Size for a Test of the Equality of Alpha Coefficients When the Number of Part-Tests is Small.” *Psychological Methods*, Vol. 4(4), pages 366-377.

Fienberg, S. 1985. *The Analysis of Cross-Classified Categorical Data*. MIT Press. Cambridge, Massachusetts. This book provides a very good introduction to the subject. It is a must for any serious student of the subject.

Finney, D. 1971. *Probit Analysis*. Cambridge University Press. New York, N.Y.

Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. Cambridge University Press. New York, New York.

Fisher, R.A. 1936. “The use of multiple measurements in taxonomic problems.” *Annuals of Eugenics*, Volume 7, Part II, 179-188. This article is famous because in it Fisher included the ‘iris data’ that is always presented when discussing discriminant analysis.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, J. L., Levin, B., Paik, M.C. 2003. *Statistical Methods for Rates and Proportions. Third Edition.* John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons. New York. This book provides a very good introduction to clinical trials. It may be a bit out of date now, but it is still very useful.

Fleming, T. R. 1982. "One-sample multiple testing procedure for Phase II clinical trials." *Biometrics*, Volume 38, pages 143-151.

Flury, B. and Riedwyl, H. 1988. *Multivariate Statistics: A Practical Approach*. Chapman and Hall. New York. This is a short, paperback text that provides lots of examples.

Flury, B. 1988. *Common Principal Components and Related Multivariate Models*. John Wiley & Sons. New York. This reference describes several advanced PCA procedures.

G

Gans. 1984. “The Search for Significance: Different Tests on the Same Data.” *The Journal of Statistical Computation and Simulation*, 1984, pages 1-21.

Gart, John J. and Nam, Jun-mo. 1988. "Approximate Interval Estimation of the Ratio in Binomial Parameters: A Review and Corrections for Skewness." *Biometrics*, Volume 44, Issue 2, 323-338.

Gart, John J. and Nam, Jun-mo. 1990. "Approximate Interval Estimation of the Difference in Binomial Parameters: Correction for Skewness and Extension to Multiple Tables." *Biometrics*, Volume 46, Issue 3, 637-643.

Gehlback, Stephen. 1988. *Interpreting the Medical Literature: Practical Epidemiology for Clinicians*. Second Edition. McGraw-Hill. New York. Telephone: (800)722-4726. The preface of this book states that its purpose is to provide the reader with a useful approach to interpreting the quantitative results given in medical literature. We reference it specifically because of its discussion of ROC curves.

Gentle, James E. 1998. *Random Number Generation and Monte Carlo Methods*. Springer. New York.

Gibbons, J. 1976. *Nonparametric Methods for Quantitative Analysis*. Holt, Rinehart and Winston. New York.

Gleason, T.C. and Staelin, R. 1975. "A proposal for handling missing data." *Psychometrika*, 40, 229-252.

Goldstein, Richard. 1989. “Power and Sample Size via MS/PC-DOS Computers,” *The American Statistician*, Volume 43, Number 4, pages 253-260. A comparative review of power analysis software that was available at that time.

Gomez, K.A. and Gomez, A. A. 1984. *Statistical Procedures for Agricultural Research*. John Wiley & Sons. New York. This reference contains worked-out examples of many complex ANOVA designs. It includes split-plot designs. We recommend it.

Graybill, Franklin. 1961. *An Introduction to Linear Statistical Models*. McGraw-Hill. New York, New York. This is an older book on the theory of linear models. It contains a few worked examples of power analysis.

Greenacre, M. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press. Orlando, Florida. This book goes through several examples. It is probably the most complete book in English on the subject.

Greenacre, Michael J. 1993. *Correspondence Analysis in Practice*. Academic Press. San Diego, CA. This book provides a self-teaching course in correspondence analysis. It is the clearest exposition on the subject that I have every seen. If you want to gain an understanding of CA, you must obtain this (paperback) book.

Griffiths, P. and Hill, I.D. 1985. *Applied Statistics Algorithms*, The Royal Statistical Society, London, England. See page 243 for ACM algorithm 291.

Gross and Clark 1975. *Survival Distributions*: Reliability Applications in Biomedical Sciences. John Wiley, New York.

Gu, X.S., and Rosenbaum, P.R. 1993. "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms," *Journal of Computational and Graphical Statistics*, Vol. 2, No. 4, pages 405-420.

Guenther, William C. 1977. "Desk Calculation of Probabilities for the Distribution of the Sample Correlation Coefficient," *The American Statistician*, Volume 31, Number 1, pages 45-48.

Guenther, William C. 1977. *Sampling Inspection in Statistical Quality Control*. Griffin's Statistical Monographs, Number 37. London.

H

Haberman, S.J. 1972. "Loglinear Fit of Contingency Tables." *Applied Statistics*. Volume 21, pages 218-225. This lists the fortran program that is used to create our LLM algorithm.

Hahn, G. J. and Meeker, W.Q. 1991. *Statistical Intervals*. John Wiley & Sons. New York.

Hambleton, R.K; Swaminathan, H; Rogers, H.J. 1991. *Fundamentals of Item Response Theory*. Sage Publications. Newbury Park, California. Phone: (805)499-0721. Provides an inexpensive, readable introduction to IRT. A good place to start.

Hamilton, L. 1991. *Regression with Graphics: A Second Course in Applied Statistics*. Brooks/Cole Publishing Company. Pacific Grove, California. This book gives a great introduction to the use of graphical analysis with regression. It is a must for any serious user of regression. It is written at an introductory level.

Hand, D.J. and Taylor, C.C. 1987. *Multivariate Analysis of Variance and Repeated Measures*. Chapman and Hall. London, England.

Hanley, J. A. and McNeil, B. J. 1982. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, 143, 29-36. April, 1982.

Hanley, J. A. and McNeil, B. J. 1983. "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases." *Radiology*, 148, 839-843. September, 1983.

Hartigan, J. 1975. *Clustering Algorithms*. John Wiley. New York. (This is the "bible" of cluster algorithms. Hartigan developed the K-means algorithm used in **NCSS**.)

Haupt, R.L. and Haupt, S.E. 1998. *Practical Genetic Algorithms*. John Wiley. New York.

Hernandez-Bermejo, B. and Sorribas, A. 2001. "Analytical Quantile Solution for the S-distribution, Random Number Generation and Statistical Data Modeling." *Biometrical Journal* 43, 1007-1025.

Hintze, J. L. and Nelson, R.D. 1998. "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician* 52, 181-184.

Hoaglin, Mosteller, and Tukey. 1985. *Exploring Data Tables, Trends, and Shapes*. John Wiley. New York.

Hoaglin, Mosteller, and Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons. New York.

Hochberg, Y. and Tamhane, A. C. 1987. *Multiple Comparison Procedures*. John Wiley & Sons. New York.

Hoerl, A.E. and Kennard, R.W. 1970. "Ridge Regression: Biased estimation for nonorthogonal problems." *Technometrics* 12, 55-82.

Hoerl, A.E. and Kennard R.W. 1976. "Ridge regression: Iterative estimation of the biasing parameter." *Communications in Statistics* A5, 77-88.

Howe, W.G. 1969. "Two-Sided Tolerance Limits for Normal Populations—Some Improvements." *Journal of the American Statistical Association*, 64, 610-620.

Hosmer, D. and Lemeshow, S. 1989. *Applied Logistic Regression*. John Wiley & Sons. New York. This book gives an advanced, in depth look at logistic regression.

Hosmer, D. and Lemeshow, S. 1999. *Applied Survival Analysis*. John Wiley & Sons. New York.

Hotelling, H. 1933. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24, 417-441, 498-520.

Hsieh, F.Y. 1989. "Sample Size Tables for Logistic Regression," *Statistics in Medicine*, Volume 8, pages 795-802. This is the article that was the basis for the sample size calculations in logistic regression in PASS 6.0. It has been superceded by the 1998 article.

Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. "A Simple Method of Sample Size Calculation for Linear and Logistic Regression," *Statistics in Medicine*, Volume 17, pages 1623-1634. The sample size calculation for logistic regression in PASS are based on this article.

Hsieh, F.Y. and Lavori, P.W. 2000. "Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates," *Controlled Clinical Trials*, Volume 21, pages 552-560. The sample size calculation for Cox regression in PASS are based on this article.

Hsu, Jason. 1996. *Multiple Comparisons: Theory and Methods.* Chapman & Hall. London. This book gives a beginning to intermediate discussion of multiple comparisons, stressing the interpretation of the various MC tests. It provides details of three popular MC situations: all pairs, versus the best, and versus a control. The power calculations used in the MC module of PASS came from this book.

I

Irizarry, R.A., et al. 2003a. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, 4, 249-264.

Irizarry, R.A., et al. 2003b. Summaries of Affymetrix GeneChip Probe Level Data. *Nucleic Acids Research*, 31, e15.

J

Jackson, J.E. 1991. *A User's Guide To Principal Components.* John Wiley & Sons. New York. This is a great book to learn about PCA from. It provides several examples and treats everything at a level that is easy to understand.

James, Mike. 1985. *Classification Algorithms*. John Wiley & Sons. New York. This is a great text on the application of discriminant analysis. It includes a simple, easy-to-understand, theoretical development as well as discussions of the application of discriminant analysis.

Jammalamadaka, S.R. and SenGupta, A. 2001. *Topics in Circular Statistics*. World Scientific. River Edge, New Jersey.

Jobson, J.D. 1992. *Applied Multivariate Data Analysis - Volume II: Categorical and Multivariate Methods*. Springer-Verlag. New York. This book is a useful reference for loglinear models and other multivariate methods. It is easy to follows and provides lots of examples.

Jolliffe, I.T. 1972. "Discarding variables in a principal component analysis, I: Artifical data." *Applied Statistics*, 21:160-173.

Johnson, N.L., Kotz, S., and Kemp, A.W. 1992. *Univariate Discrete Distributions, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions Volume 1, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1995. *Continuous Univariate Distributions Volume 2, Second Edition*. John Wiley & Sons. New York.

Jolliffe, I.T. 1986. *Principal Component Analysis*. Springer-Verlag. New York. This book provides an easy-reading introduction to PCA. It goes through several examples.

Julious, Steven A. 2004. "Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data." *Statistics in Medicine*, 23:1921-1986.

Juran, J.M. 1979. *Quality Control Handbook*. McGraw-Hill. New York.

K

Kaiser, H.F. 1960. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement*. 20:141-151.

Kalbfleisch, J.D. and Prentice, R.L. 1980. *The Statistical Analysis of Failure Time Data*. John Wiley, New York.

Karian, Z.A and Dudewicz, E.J. 2000. *Fitting Statistical Distributions.* CRC Press, New York.

Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data*. John Wiley. New York. This book gives an excellent introduction to cluster analysis. It treats the forming of the distance matrix and several different types of cluster methods, including fuzzy. All this is done at an elementary level so that users at all levels can gain from it.

Kay, S.M. 1988. *Modern Spectral Estimation*. Prentice-Hall: Englewood Cliffs, New Jersey. A very technical book on spectral theory.

Kendall,M. and Ord, J.K. 1990. *Time Series.* Oxford University Press. New York. This is theoretical introduction to time series analysis that is very readable.

Kendall,M. and Stuart, A. 1987. *Kendall's Advanced Theory of Statistics. Volume 1: Distribution Theory.* Oxford University Press. New York. This is a fine math-stat book for graduate students in statistics. We reference it because it includes formulas that are used in the program.

Kenward, M. G. and Roger, J. H. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics,* 53, pages 983-997.

Keppel, Geoffrey. 1991. *Design and Analysis - A Researcher's Handbook.* Prentice Hall. Englewood Cliffs, New Jersey. This is a very readable primer on the topic of analysis of variance. Recommended for those who want the straight scoop with a few, well-chosen examples.

Kirk, Roger E. 1982. *Experimental Design: Procedures for the Behavioral Sciences.* Brooks/Cole. Pacific Grove, California. This is a respected reference on experimental design and analysis of variance.

Klein, J.P. and Moeschberger, M.L.. 1997. *Survival Analysis.* Springer-Verlag. New York. This book provides a comprehensive look at the subject complete with formulas, examples, and lots of useful comments. It includes all the more recent developments in this field. I recommend it.

Koch, G.G.; Atkinson, S.S.; Stokes, M.E. 1986. *Encyclopedia of Statistical Sciences.* Volume 7. John Wiley. New York. Edited by Samuel Kotz and Norman Johnson. The article on Poisson Regression provides a very good summary of the subject.

Kotz and Johnson. 1993. *Process Capability Indices.* Chapman & Hall. New York. This book gives a detailed account of the capability indices used in SPC work. 207 pages.

Kraemer, H. C. and **Thiemann, S.** 1987. *How Many Subjects*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Kruskal, J. 1964. “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.” *Psychometrika* 29, pages 1-27, 115-129. This article presents the algorithm on which the non-metric algorithm used in NCSS is based.

Kruskal, J. and Wish, M. 1978. *Multidimensional Scaling*. Sage Publications. Beverly Hills, CA. This is a well-written monograph by two of the early pioneers of MDS. We suggest it to all serious students of MDS.

Kuehl, R.O. 2000. *Design of Experiment: Statistical Principles of Research Design and Analysis, 2nd Edition.* Brooks/Cole. Pacific Grove, California. This is a good graduate level text on experimental design with many examples.

L

Lachenbruch, P.A. 1975. *Discriminant Analysis.* Hafner Press. New York. This is an in-depth treatment of the subject. It covers a lot of territory, but has few examples.

Lachin, John M. 2000. *Biostatistical Methods.* John Wiley & Sons. New York. This is a graduate-level methods book that deals with statistical methods that are of interest to biostatisticians such as odds ratios, relative risks, regression analysis, case-control studies, and so on.

Lachin, John M. and **Foulkes, Mary A. 1986.** “Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification,” *Biometrics*, Volume 42, September, pages 507-516.

Lan, K.K.G. and DeMets, D.L. 1983. “Discrete sequential boundaries for clinical trials.” *Biometrika*, 70, pages 659-663.

Lan, K.K.G. and Zucker, D.M. 1993. “Sequential monitoring of clinical trials: the role of information and Brownian motion.” *Statistics in Medicine*, 12, pages 753-765.

Lance, G.N. and Williams, W.T. 1967. “A general theory of classificatory sorting strategies. I. Hierarchical systems.” *Comput. J.* 9, pages 373-380.

Lance, G.N. and Williams, W.T. 1967. “Mixed-data classificatory programs I. Agglomerative systems.” *Aust.Comput. J.* 1, pages 15-20.

Lawless, J.F. 1982. *Statistical Models and Methods for Lifetime Data*. John Wiley, New York.

Lawson, John. 1987. *Basic Industrial Experimental Design Strategies*. Center for Statistical Research at Brigham Young University. Provo, Utah. 84602. This is a manuscript used by Dr. Lawson in courses and workshops that he provides to industrial engineers. It is the basis for many of our experimental design procedures.

Lebart, Morineau, and Warwick. 1984. *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons. This book devotes a large percentage of its discussion to correspondence analysis.

Lee, E.T. 1974. "A Computer Program for Linear Logistic Regression Analysis" in *Computer Programs in Biomedicine,* Volume 4, pages 80-92.

Lee, E.T. 1980. *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications. Belmont, California.

Lee, E.T. 1992. *Statistical Methods for Survival Data Analysis*. Second Edition. John Wiley & Sons. New York. This book provides a very readable introduction to survival analysis techniques.

Lee, S. K. 1977. "On the Asymptotic Variances of u Terms in Loglinear Models of Multidimensional Contingency Tables." *Journal of the American Statistical Association*. Volume 72 (June, 1977), page 412. This article describes methods for computing standard errors that are used in the LLM section of this program.

Lenth, Russell V. 1987. "Algorithm AS 226: Computing Noncentral Beta Probabilities," *Applied Statistics*, Volume 36, pages 241-244.

Lenth, Russell V. 1989. "Algorithm AS 243: Cumulative Distribution Function of the Non-central t Distribution," *Applied Statistics*, Volume 38, pages 185-189.

Lesaffre, E. and Albert, A. 1989. "Multiple-group Logistic Regression Diagnostics" *Applied Statistics*, Volume 38, pages 425-440. See also Pregibon 1981.

Levene, H. 1960. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al., eds. Stanford University Press, Stanford Calif., pp. 278-292.

Lewis, J.A. 1999. "Statistical principles for clinical trials (ICH E9) an introductory note on an international guideline." *Statistics in Medicine*, 18, pages 1903-1942.

Lipsey, Mark W. 1990. *Design Sensitivity Statistical Power for Experimental Research*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Little, R. and Rubin, D. 1987. *Statistical Analysis with Missing Data*. John Wiley & Sons. New York. This book is completely devoted to dealing with missing values. It gives a complete treatment of using the EM algorithm to estimate the covariance matrix.

Little, R. C. et al. 2006. *SAS for Mixed Models – Second Edition.* SAS Institute Inc., Cary, North Carolina.

Liu, H. and Wu, T. 2005. "Sample Size Calculation and Power Analysis of Time-Averaged Difference," *Journal of Modern Applied Statistical Methods*, Vol. 4, No. 2, pages 434-445.

Lu, Y. and Bean, J.A. 1995. "On the sample size for one-sided equivalence of sensitivities based upon McNemar's test," *Statistics in Medicine*, Volume 14, pages 1831-1839.

Lui, J., Hsueh, H., Hsieh, E., and Chen, J.J. 2002. "Tests for equivalence or non-inferiority for paired binary data," *Statistics in Medicine*, Volume 21, pages 231-245.

Lloyd, D.K. and Lipow, M. 1991. *Reliability: Management, Methods, and Mathematics*. ASQC Quality Press. Milwaukee, Wisconsin.

Locke, C.S. 1984. "An exact confidence interval for untransformed data for the ratio of two formulation means," *J. Pharmacokinet. Biopharm.*, Volume 12, pages 649-655.

Lockhart, R. A. & Stephens, M. A. 1985. "Tests of fit for the von Mises distribution." *Biometrika* 72, pages 647-652.

M

Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. *Sample Size Tables for Clinical Studies, 2nd Edition*. Blackwell Science. Malden, Mass. A very good & easy to read book on determining appropriate sample sizes in many situations.

Makridakis, S. and Wheelwright, S.C. 1978. *Iterative Forecasting*. Holden-Day.: San Francisco, California. This is a very good book for the layman since it includes several detailed examples. It is written for a person with a minimum amount of mathematical background.

Manly, B.F.J. 1986. *Multivariate Statistical Methods - A Primer*. Chapman and Hall. New York. This nice little paperback provides a simplified introduction to many multivariate techniques, including MDS.

Mardia, K.V. and Jupp, P.E. 2000. *Directional Statistics*. John Wiley & Sons. New York.

Marple, S.L. 1987. *Digital Spectral Analysis with Applications*. Prentice-Hall: Englewood Cliffs, New Jersey. A technical book about spectral analysis.

Martinez and Iglewicz. 1981. “A test for departure from normality based on a biweight estimator of scale.” *Biometrika*, 68, 331-333).

Marubini, E. and **Valsecchi, M.G.** 1996. *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley: New York, New York.

Mather, Paul. 1976. *Computational Methods of Multivariate Analysis in Physical Geography*. John Wiley & Sons. This is a great book for getting the details on several multivariate procedures. It was written for non-statisticians. It is especially useful in its presentation of cluster analysis. Unfortunately, it is out-of-print. You will have to look for it in a university library (it is worth the hunt).

Matsumoto, M. and Nishimura,T. 1998. “Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator” *ACM Trans. On Modeling and Computer Simulations*.

Mauchly, J.W. 1940. “Significance test for sphericity of a normal n-variate distribution.” *Annals of Mathematical Statistics*, 11: 204-209

McCabe, G.P. 1984. “Principal variables.” *Technometrics*, 26, 137-144.

McClish, D.K. 1989. “Analyzing a Portion of the ROC Curve.” *Medical Decision Making*, 9: 190-195

McHenry, Claude. 1978. "Multivariate subset selection." *Journal of the Royal Statistical Society, Series C*. Volume 27, No. 23, pages 291-296.

McNeil, D.R. 1977. *Interactive Data Analysis*. John Wiley & Sons. New York.

Mendenhall, W. 1968. *Introduction to Linear Models and the Design and Analysis of Experiments*. Wadsworth. Belmont, Calif.

Metz, C.E. 1978. “Basic principles of ROC analysis.” *Seminars in Nuclear Medicine*, Volume 8, No. 4, pages 283-298.

Miettinen, O.S. and Nurminen, M. 1985. “Comparative analysis of two rates.” *Statistics in Medicine* 4: 213-226.

Milliken, G.A. and Johnson, D.E. 1984. *Analysis of Messy Data, Volume 1*. Van Nostrand Rienhold. New York, NY.

Milne, P. 1987. *Computer Graphics for Surveying.* E. & F. N. Spon, 29 West 35th St., NY, NY 10001

Montgomery, Douglas. 1984. *Design and Analysis of Experiments*. John Wiley & Sons, New York. A textbook covering a broad range of experimental design methods. The book is not limited to industrial investigations, but gives a much more general overview of experimental design methodology.

Montgomery, Douglas and Peck. 1992. *Introduction to Linear Regression Analysis*. A very good book on this topic.

Montgomery, Douglas C. 1991. *Introduction to Statistical Quality Control.* Second edition. John Wiley & Sons. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 700 pages.

Moore, D. S. and McCabe, G. P. 1999. *Introduction to the Practice of Statistics*. W. H. Freeman and Company. New York.

Mosteller, F. and Tukey, J.W. 1977. *Data Analysis and Regression*. Addison-Wesley. Menlo Park, California. This book should be read by all serious users of regression analysis. Although the terminology is a little different, this book will give you a fresh look at the whole subject.

Motulsky, Harvey. 1995. *Intuitive Biostatistics.* Oxford University Press. New York, New York. This is a wonderful book for those who want to understand the basic concepts of statistical testing. The author presents a very readable coverage of the most popular biostatistics tests. If you have forgotten how to interpret the various statistical tests, get this book!

Moura, Eduardo C. 1991. *How To Determine Sample Size And Estimate Failure Rate in Life Testing*. ASQC Quality Press. Milwaukee, Wisconsin.

Mueller, K. E., and Barton, C. N. 1989. “Approximate Power for Repeated-Measures ANOVA Lacking Sphericity.” *Journal of the American Statistical Association*, Volume 84, No. 406, pages 549-555.

Mueller, K. E., LaVange, L.E., Ramey, S.L., and Ramey, C.T. 1992. “Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications.” *Journal of the American Statistical Association*, Volume 87, No. 420, pages 1209-1226.

Mukerjee, H., Robertson, T., and Wright, F.T. 1987. “Comparison of Several Treatments With a Control Using Multiple Contrasts.” *Journal of the American Statistical Association*, Volume 82, No. 399, pages 902-910.

Muller, K. E. and Stewart, P.W. 2006. *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley & Sons Inc. Hoboken, New Jersey.

Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston, Massachusetts. This is one of the bibles on the topic of regression analysis.

N

Naef, F. et al. 2002. “Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays,” *Genome Biol.*, 3, RESEARCH0018.

Nam, Jun-mo. 1992. “Sample Size Determination for Case-Control Studies and the Comparison of Stratified and Unstratified Analyses,” *Biometrics*, Volume 48, pages 389-395.

Nam, Jun-mo. 1997. “Establishing equivalence of two treatments and sample size requirements in matched-pairs design,” *Biometrics*, Volume 53, pages 1422-1430.

Nam, J-m. and Blackwelder, W.C. 2002. “Analysis of the ratio of marginal probabilities in a matched-pair setting,” *Statistics in Medicine*, Volume 21, pages 689-699.

Nash, J. C. 1987. *Nonlinear Parameter Estimation*. Marcel Dekker, Inc. New York, NY.

Nash, J.C. 1979. *Compact Numerical Methods for Computers*. John Wiley & Sons. New York, NY.

Nel, D.G. and van der Merwe, C.A. 1986. “A solution to the multivariate Behrens-Fisher problem.” *Communications in Statistics—Series A, Theory and Methods*, 15, pages 3719-3735.

Nelson, W.B. 1982. *Applied Life Data Analysis*. John Wiley, New York.

Nelson, W.B. 1990. *Accelerated Testing*. John Wiley, New York.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. 1996. *Applied Linear Statistical Models*. Richard D. Irwin, Inc. Chicago, Illinois. This mammoth book covers regression analysis and analysis of variance thoroughly and in great detail. We recommend it.

Neter, J., Wasserman, W., and Kutner, M. 1983. *Applied Linear Regression Models*. Richard D. Irwin, Inc. Chicago, Illinois. This book provides you with a complete introduction to the methods of regression analysis. We suggest it to non-statisticians as a great reference tool.

Newcombe, Robert G. 1998a. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods." *Statistics in Medicine*, Volume 17, 857-872.

Newcombe, Robert G. 1998b. "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods." *Statistics in Medicine*, Volume 17, 873-890.

Newcombe, Robert G. 1998c. "Improved Confidence Intervals for the Difference Between Binomial Proportions Based on Paired Data." *Statistics in Medicine*, Volume 17, 2635-2650.

Newton, H.J. 1988. *TIMESLAB: A Time Series Analysis Laboratory*. Wadsworth & Brooks/Cole: Pacific Grove, California. This book is loaded with theoretical information about time series analysis. It includes software designed by Dr. Newton for performing advanced time series and spectral analysis. The book requires a strong math and statistical background.

O

O'Brien, P.C. and Fleming, T.R. 1979. "A multiple testing procedure for clinical trials." *Biometrics*, 35, pages 549-556.

O'Brien, R.G. and Kaiser, M.K. 1985. "MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer." *Psychological Bulletin*, 97, pages 316-333.

Obuchowski, N. 1998. "Sample Size Calculations in Studies of Test Accuracy." *Statistical Methods in Medical Research*, 7, pages 371-392.

Obuchowski, N. and McClish, D. 1997. "Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices." *Statistics in Medicine*, 16, pages 1529-1542.

Odeh, R.E. and Fox, M. 1991. *Sample Size Choice*. Marcel Dekker, Inc. New York, NY.

O'Neill and Wetherill. 1971 "The Present State of Multiple Comparison Methods," *The Journal of the Royal Statistical Society*, Series B, vol.33, 218-250).

Orloci, L. & Kenkel, N. 1985. *Introduction to Data Analysis*. International Co-operative Publishing House. Fairland, Maryland. This book was written for ecologists. It contains samples and BASIC programs of many statistical procedures. It has one brief chapter on MDS, and it includes a non-metric MDS algorithm.

Ostle, B. 1988. *Statistics in Research. Fourth Edition*. Iowa State Press. Ames, Iowa. A comprehension book on statistical methods.

Ott, L. 1977. *An Introduction to Statistical Methods and Data Analysis*. Wadsworth. Belmont, Calif. Use the second edition.

Ott, L. 1984. *An Introduction to Statistical Methods and Data Analysis, Second Edition*. Wadsworth. Belmont, Calif. This is a complete methods text. Regression analysis is the focus of five or six chapters. It stresses the interpretation of the statistics rather than the calculation, hence it provides a good companion to a statistical program like ours.

Owen, Donald B. 1956. "Tables for Computing Bivariate Normal Probabilities," *Annals of Mathematical Statistics*, Volume 27, pages 1075-1090.

Owen, Donald B. 1965. "A Special Case of a Bivariate Non-Central t-Distribution," *Biometrika*, Volume 52, pages 437-446.

P

Pandit, S.M. and Wu, S.M. 1983. *Time Series and System Analysis with Applications*. John Wiley and Sons. New York. This book provides an alternative to the Box-Jenkins approach for dealing with ARMA models. We used this approach in developing our automatic ARMA module.

Parmar, M.K.B. and Machin, D. 1995. *Survival Analysis*. John Wiley and Sons. New York.

Parmar, M.K.B., Torri, V., and Steart, L. 1998. "Extracting Summary Statistics to Perform Meta-Analyses of the Published Literature for Survival Endpoints." *Statistics in Medicine* 17, 2815-2834.

Pearson, K. 1901. "On lines and planes of closest fit to a system of points in space." *Philosophical Magazine* 2, 557-572.

Pan, Z. and Kupper, L. 1999. "Sample Size Determination for Multiple Comparison Studies Treating Confidence Interval Width as Random." *Statistics in Medicine* 18, 1475-1488.

Pedhazur, E.L. and Schmelkin, L.P. 1991. *Measurement, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. This mammoth book (over 800 pages) covers multivariate analysis, regression analysis, experimental design, analysis of variance, and much more. It provides annotated output from SPSS and SAS which is also useful to our users. The text is emphasizes the social sciences. It provides a "how-to," rather than a theoretical, discussion. Its chapters on factor analysis are especially informative.

Phillips, Kem F. 1990. “Power of the Two One-Sided Tests Procedure in Bioequivalence,” *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 18, No. 2, pages 137-144.

Pocock, S.J. 1977. “Group sequential methods in the design and analysis of clinical trials.” *Biometrika*, 64, pages 191-199.

Press, S. J. and Wilson, S. 1978. “Choosing Between Logistic Regression and Discriminant Analysis.” *Journal of the American Statistical Association*, Volume 73, Number 364, Pages 699-705. This article details the reasons why logistic regression should be the preferred technique.

Press, William H. 1986. *Numerical Recipes*, Cambridge University Press, New York, New York.

Pregibon, Daryl. 1981. “Logistic Regression Diagnostics.” *Annals of Statistics*, Volume 9, Pages 705-725. This article details the extensions of the usual regression diagnostics to the case of logistic regression. These results were extended to multiple-group logistic regression in Lesaffre and Albert (1989).

Price, K., Storn R., and Lampinen, J. 2005. *Differential Evolution – A Practical Approach to Global Optimization.* Springer. Berlin, Germany.

Prihoda, Tom. 1983. “Convenient Power Analysis For Complex Analysis of Variance Models.” *Poster Session of the American Statistical Association Joint Statistical Meetings*, August 15-18, 1983, Toronto, Canada. Tom is currently at the University of Texas Health Science Center. This article includes FORTRAN code for performing power analysis.

R

Ramsey, Philip H. 1978 “Power Differences Between Pairwise Multiple Comparisons,” *JASA*, vol. 73, no. 363, pages 479-485.

Rao, C.R. , Mitra, S.K., & Matthai, A. 1966. *Formulae and Tables for Statistical Work.* Statistical Publishing Society, Indian Statistical Institute, Calcutta, India.

Ratkowsky, David A. 1989. *Handbook of Nonlinear Regression Models*. Marcel Dekker. New York. A good, but technical, discussion of various nonlinear regression models.

Rawlings John O. 1988. *Applied Regression Analysis: A Research Tool*. Wadsworth. Belmont, California. This is a readable book on regression analysis. It provides a thorough discourse on the subject.

Reboussin, D.M., DeMets, D.L., Kim, K, and Lan, K.K.G. 1992. “Programs for computing group sequential boundaries using the Lan-DeMets Method.” Technical Report 60, Department of Biostatistics, University of Wisconsin-Madison.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. John Wiley. New York, New York. This book provides a comprehensive mixture of theoretical and applied results in multivariate analysis. My evaluation may be biased since Al Rencher took me fishing when I was his student.

Robins, Greenland, and Breslow. 1986. “A General Estimator for the Variance of the Mantel-Haenszel Odds Ratio,” *American Journal of Epidemiology*, vol.42, pages 719-723.

Robins, Breslow, and Greenland. 1986. “Estimators of the Mantel-Haenszel variance consisten in both sparse data and large-strata limiting models,” *Biometrics*, vol. 42, pages 311-323.

Rosenbaum, P.R. 1989. “Optimal Matching for Observational Studies,” *Journal of the American Statistical Association*, vol. 84, no. 408, pages 1024-1032.

Rosenbaum, P.R., and Rubin, D.B. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, vol. 70, pages 41-55.

Rosenbaum, P.R., and Rubin, D.B. 1984. “Reducing bias in observational studies using subclassification on the propensity score,” *Journal of the American Statistical Association*, vol. 79, pages 516-524.

Rosenbaum, P.R., and Rubin, D.B. 1985a. “Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score,” *American Statistician*, vol. 39, pages 33-38.

Rosenbaum, P.R., and Rubin, D.B. 1985b. “The Bias Due to Incomplete Matching,” *Biometrics*, vol. 41, pages 106-116.

Ryan, Thomas P. 1989. *Statistical Methods for Quality Improvement*. John Wiley & Sons. New York. This is a comprehensive treatment of SPC including control charts, process capability, and experimental design. It provides many rules-of-thumb and discusses many non-standard situations. This is a very good ‘operators manual’ type of book. 446 pages.

Ryan, Thomas P. 1997. *Modern Regression Methods*. John Wiley & Sons. New York. This is a comprehensive treatment of regression analysis. The author often deals with practical issues that are left out of other texts.

S

Sahai, Hardeo & Khurshid, Anwer. 1995. *Statistics in Epidemiology*. CRC Press. Boca Raton, Florida.

Schiffman, Reynolds, & Young. 1981. *Introduction to Multidimensional Scaling*. Academic Press. Orlando, Florida. This book goes through several examples.

Schilling, Edward. 1982. *Acceptance Sampling in Quality Control*. Marcel-Dekker. New York.

Schlesselman, Jim. 1981. *Case-Control Studies*. Oxford University Press. New York. This presents a complete overview of case-control studies. It was our primary source for the Mantel-Haenszel test.

Schmee and Hahn. November, 1979. "A Simple Method for Regression Analysis." *Technometrics*, Volume 21, Number 4, pages 417-432.

Schoenfeld, David A. 1983. “Sample-Size Formula for the Proportional-Hazards Regression Model” *Biometrics*, Volume 39, pages 499-503.

Schoenfeld, David A. and **Richter, Jane R.** 1982**.** “Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint,” *Biometrics,* March 1982, Volume 38, pages 163-170.

Schork, M. and Williams, G. 1980. "Number of Observations Required for the Comparison of Two Correlated Proportions." *Communications in Statistics-Simula. Computa.,* B9(4), 349-357.

Schuirmann, Donald. 1981**.** “On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval,” *Biometrics,* Volume 37, pages 617.

Schuirmann, Donald. 1987**.** “A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability,” *Journal of Pharmacokinetics and Biopharmaceutics,* Volume 15, Number 6, pages 657-680.

Seber, G.A.F. 1984. *Multivariate Observations*. John Wiley & Sons. New York. (This book is an encyclopedia of multivariate techniques. It emphasizes the mathematical details of each technique and provides a complete set of references. It will only be useful to those comfortable with reading mathematical equations based on matrices.)

Seber, G.A.F. and Wild, C.J. 1989. *Nonlinear Regression*. John Wiley & Sons. New York. This book is an encyclopedia of nonlinear regression.

Senn, Stephen. 1993. *Cross-over Trials in Clinical Research*. John Wiley & Sons. New York.

Senn, Stephen. 2002. *Cross-over Trials in Clinical Research*. Second Edition. John Wiley & Sons. New York.

Shapiro, S.S. and Wilk, M.B. 1965 "An analysis of Variance test for normality." *Biometrika*, Volume 52, pages 591-611.

Shuster, Jonathan J. 1990. *CRC Handbook of Sample Size Guidelines for Clinical Trials.* CRC Press, Boca Raton, Florida. This is an expensive book ($300) of tables for running log-rank tests. It is well documented, but at this price it better be.

Signorini, David. 1991. “Sample size for Poisson regression,” *Biometrika,* Volume 78, 2, pages 446-450.

Simon, Richard. “Optimal Two-Stage Designs for Phase II Clinical Trials,” *Controlled Clinical Trials,* 1989, Volume 10, pages 1-10.

Snedecor, G. and Cochran, Wm. 1972. *Statistical Methods*. The Iowa State University Press. Ames, Iowa.

Sorribas, A., March, J., and Trujillano, J. 2002. “A new parametric method based on S-distributions for computing receiver operating characteristic curves for continuous diagnostic tests.” *Statistics in Medicine* 21, 1213-1235.

Spath, H. 1985. *Cluster Dissection and Analysis.* Halsted Press. New York. (This book contains a detailed discussion of clustering techniques for large data sets. It contains some heavy mathematical notation.)

Speed, T.P. (editor). 2003. *Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall/CRC. Boca Raton, Florida.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., and Song, F. 2000. *Methods for Meta-Analysis in Medical Research.* John Wiley & Sons. New York.

Swets, John A. 1996. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics - Collected Papers.* Lawrence Erlbaum Associates. Mahway, New Jersey.

T

Tabachnick, B. and Fidell, L. 1989. *Using Multivariate Statistics.* Harper Collins. 10 East 53d Street, NY, NY 10022. This is an extremely useful text on multivariate techniques. It presents computer printouts and discussion from several popular programs. It provides checklists for each procedure as well as sample written reports. I strongly encourage you to obtain this book!

Tango, Toshiro. 1998. "Equivalence Test and Confidence Interval for the Difference in Proportions for the Paired-Sample Design." *Statistics in Medicine*, Volume 17, 891-908.

Therneau, T.M. and Grambsch, P.M. 2000. *Modeling Survival Data*. Springer: New York, New York. A the time of the writing of the Cox regression procedure, this book provides a thorough, up-to-date discussion of this procedure as well as many extensions to it. Recommended, especially to those with at least a masters in statistics.

Thomopoulos, N.T. 1980. *Applied Forecasting Methods*. Prentice-Hall: Englewood Cliffs, New Jersey. This book contains a very good presentation of the classical forecasting methods discussed in chapter two.

Thompson, Simon G. 1998. *Encyclopedia of Biostatistics, Volume 4*. John Wiley & Sons. New York. Article on Meta-Analysis on pages 2570-2579.

Tiku, M. L. 1965. "Laguerre Series Forms of Non-Central X^2 and F Distributions," *Biometrika*, Volume 42, pages 415-427.

Torgenson, W.S. 1952. "Multidimensional scaling. I. Theory and method." *Psychometrika* 17, 401-419. This is one of the first articles on MDS. There have been many advances, but this article presents many insights into the application of the technique. It describes the algorithm on which the metric solution used in this program is based.

Tubert-Bitter, P., Manfredi,R., Lellouch, J., Begaud, B. 2000. "Sample size calculations for risk equivalence testing in pharmacoepidemiology." *Journal of Clinical Epidemiology* 53, 1268-1274.

Tukey, J.W. and McLaughlin, D.H. 1963. "Less Vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization." *Sankhya, Series A* 25, 331-352.

Tukey, J.W. 1977. *Exploratory Data Analysis.* Addison-Wesley Publishing Company. Reading, Mass.

U

Upton, G.J.G. 1982."A Comparison of Alternative Tests for the 2 x 2 Comparative Trial.", *Journal of the Royal Statistical Society,* Series A,, Volume 145, pages 86-105.

Upton, G.J.G. and Fingleton, B. 1989. *Spatial Data Analysis by Example: Categorical and Directional Data. Volume 2.* John Wiley & Sons. New York.

V

Velicer, W.F. 1976. "Determining the number of components from the matrix of partial correlations." *Psychometrika*, 41, 321-327.

Velleman, Hoaglin. 1981. *ABC's of Exploratory Data Analysis.* Duxbury Press, Boston, Massachusetts.

Voit, E.O. 1992. "The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions." *Biometrical J.* 34, 855-878.

Voit, E.O. 2000. "A Maximum Likelihood Estimator for Shape Parameters of S-Distributions." *Biometrical J.* 42, 471-479.

Voit, E.O. and Schwacke, L. 1998. "Scalability properties of the S-distribution." *Biometrical J.* 40, 665-684.

Voit, E.O. and Yu, S. 1994. "The S-distribution. Approximation of discrete distributions." *Biometrical J.* 36, 205-219.

W

Walter, S.D., Eliasziw, M., and Donner, A. 1998. "Sample Size and Optimal Designs For Reliability Studies." *Statistics in Medicine*, 17, 101-110.

Welch, B.L. 1938. "The significance of the difference between two means when the population variances are unequal." *Biometrika*, 29, 350-362.

Welch, B.L. 1947. "The Generalization of "Student's" Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Welch, B.L. 1949. "Further Note on Mrs. Aspin's Tables and on Certain Approximations to the Tabled Function," *Biometrika*, 36, 293-296.

Westfall, P. et al. 1999. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc. Cary, North Carolina.

Westgard, J.O. 1981. "A Multi-Rule Shewhart Chart for Quality Control in Clinical Chemistry," *Clinical Chemistry*, Volume 27, No. 3, pages 493-501. (This paper is available online at the www.westgard.com).

Westlake, W.J. 1981. "Bioequivalence testing—a need to rethink," *Biometrics*, Volume 37, pages 591-593.

Whittemore, Alice. 1981. "Sample Size for Logistic Regression with Small Response Probability," *Journal of the American Statistical Association*, Volume 76, pages 27-32.

Wickens, T.D. 1989. *Multiway Contingency Tables Analysis for the Social Sciences.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. A thorough book on the subject. Discusses loglinear models in depth.

Wilson, E.B.. 1927. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, Volume 22, pages 209-212. This article discusses the 'score' method that has become popular when dealing with proportions.

Winer, B.J. 1991**.** *Statistical Principles in Experimental Design (Third Edition)*. McGraw-Hill. New York, NY. A very complete analysis of variance book.

Wit, E., and McClure, J. 2004. *Statistics for Microarrays*. John Wiley & Sons Ltd, Chichester, West Sussex, England.

Wolfinger, R., Tobias, R. and Sall, J. 1994. "Computing Gaussian likelihoods and their derivatives for general linear mixed models," *SIAM Journal of Scientific Computing*, 15, no.6, pages 1294-1310.

Woolson, R.F., Bean, J.A., and Rojas, P.B. 1986. "Sample Size for Case-Control Studies Using Cochran's Statistic," *Biometrics*, Volume 42, pages 927-932.

Y

Yuen, K.K. and Dixon, W. J. 1973. “The approximate behavior and performance of the two-sample trimmed t,” *Biometrika*, Volume 60, pages 369-374.

Yuen, K.K. 1974. “The two-sample trimmed t for unequal population variances,” *Biometrika*, Volume 61, pages 165-170.

Z

Zar, Jerrold H. 1984**.** *Biostatistical Analysis (Second Edition)*. Prentice-Hall. Englewood Cliffs, New Jersey. This introductory book presents a nice blend of theory, methods, and examples for a long list of topics of interest in biostatistical work.

Zhou, X., Obuchowski, N., McClish, D. 2002**.** *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc. New York, New York. This is a great book on the designing and analyzing diagnostic tests. It is especially useful for its presentation of ROC curves.

Chapter Index

Index