Affymetrix® Tiling Analysis Software Version 1.1 – User Guide

PN 702323 Rev. 1

For research use only.
Not for use in diagnostic procedures.

Trademarks

Limited License Notice

Patents

Copyright

Table of Contents

Welcome

Welcome

Welcome to the Affymetrix® Tiling Analysis Software (TAS) v1.1 User Guide. TAS provides analysis capabilities for the Affymetrix GeneChip® Tiling Array.

Introduction

Analysis functions provided within the TAS application include:

- Analyzing feature-intensity data stored in CEL files to produce signal and p-values for each interrogated genomic probe position
- Computation of genomic intervals based on computed signal and p-values
- Computation of summary statistics
- Visualizations for assessing the quality of array data

Analyses produced using TAS can be imported into applications such as the Integrated Genome Browser (IGB) or the UCSC (University of California Santa Cruz) Genome Browser for visualization against genomic annotations.

Display Features

The main TAS window contains two major sections:

- Data stored in files (top half of the software window)
- Status messages for workflow (bottom half of the software window)

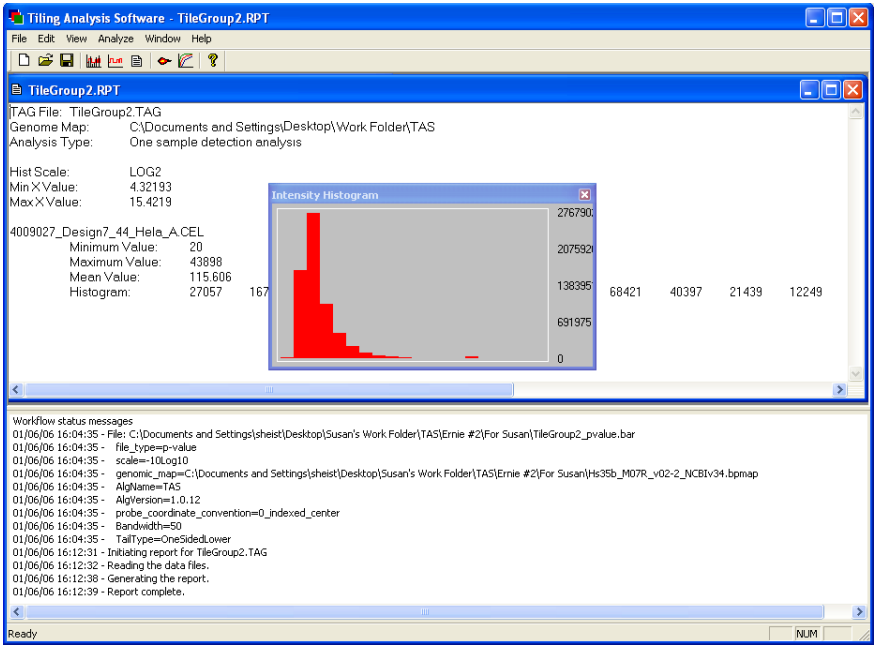

Figure 1.1
TAS main window

Data Files

The TAS application manages several file types that store data from analysis results, reports and graphs: CEL, BPMAP, TAG, ROC, MVA, BAR, BED, RPT, and GFF. See *Appendix A, TAS File Types* for detailed descriptions of each file type.

Files that store analyses and quality control results such as ROC plots, MVA plots and RPT reports are available for onscreen display. TAS is capable of displaying these files simultaneously (see Figure 1.2).

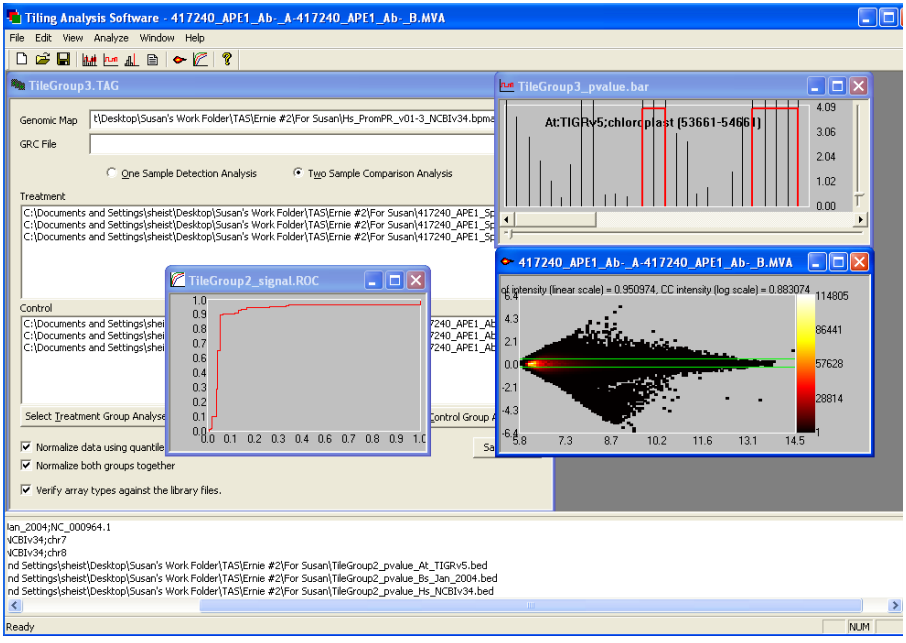

Figure 1.2
TAS window with multiple files types displayed (ROC plot, BAR plot, MVA plot, and TAG file)

Workflow

The Affymetrix® Tiling Analysis Software (TAS) is used to analyze Affymetrix GeneChip® Tiling Array intensity data. The workflow displayed below shows the main steps for analyzing intensity data with TAS.

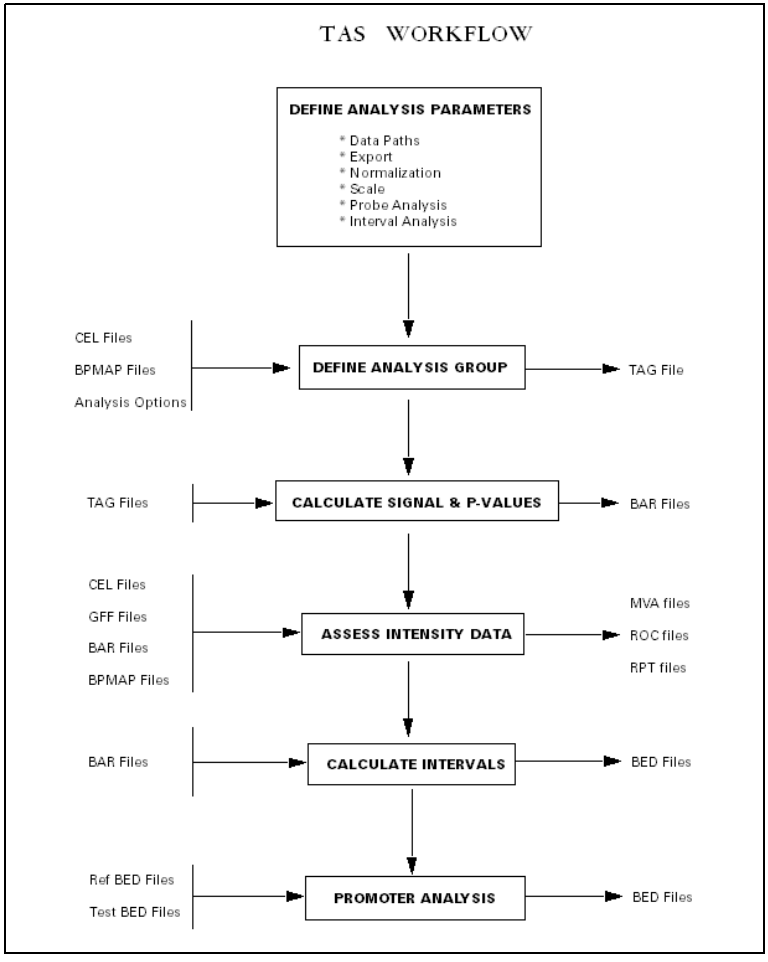

Figure 1.3
TAS workflow

About this Manual

This manual presents information about using the TAS application in the following chapters and appendices:

- *Chapter 1, Setting Analysis Parameters*
- *Chapter 2, Defining The Analysis Group*
- *Chapter 3, Analyzing Intensity Data*

The appendices include additional information and/or reference information:

- *Appendix A, TAS File Types*
- *Appendix B, Algorithms and Parameters*

CONVENTIONS USED IN THIS GUIDE

This manual provides a detailed outline for all tasks associated with the Affymetrix® TAS application. Various conventions are used throughout the manual to help illustrate the procedures described. Explanations of these conventions are provided below.

Steps

Instructions for procedures are written in a numbered step format. Immediately following the step number is the action to be performed. Following the response, additional information pertaining to the step may be found and is presented in paragraph format. For example:

1. Click **Yes** to continue.

The Delete task proceeds.
In the lower right pane the status is displayed.

To view more information pertaining to the delete task, right-click **Delete** and select **View Task Log** from the shortcut menu.

Font Styles

Bold fonts indicate names of commands, buttons, options or titles within a dialog box. When asked to enter specific information, such input opens in italics within the procedure being outlined.
For example:

1. Click the **Find** toolbar button  ; or

   Select **Edit** → **Find** from the menu bar.

   The Find dialog box opens.
2. Enter EC_RNA.cel in the **Find what** box, then click **Find Next** to view the first search result.
3. Continue to click **Find Next** to view each successive search result.

Screen Captures

The steps outlining procedures are frequently supplemented with screen captures to further illustrate the instructions given.

NOTE **The screen captures depicted in this manual may not exactly match the windows displayed on your screen.**

Additional Comments

Throughout the manual, text and procedures are occasionally accompanied by special notes. These additional comments and their meanings are described below.

TIP **Information presented in Tips provide helpful advice or shortcuts for completing a task.**

NOTE **The Note format presents supplemental information pertaining to the text or procedure being outlined.**

IMPORTANT !

The Important format presents important information that may affect the accuracy of your results.

CAUTION

Caution notes advise you that the consequence(s) of an action may be irreversible and/or result in lost data.

WARNING ▲

Warnings alert you to situations where physical harm to person or damage to hardware is possible.

Technical Support

SOFTWARE SUPPORT POLICY

TAS is an Affymetrix Tool. Affymetrix Tools are freely available, rapidly developed, and supported, but they are released AS-IS without warranty. Furthermore, while Affymetrix Tools are tested, they are not formally validated. In providing support, Affymetrix will provide technical information, training materials, and advice to users; but due to the rapid evolution and development of these tools, Affymetrix may not be able to answer and resolve all support inquiries or problems observed.

CONTACT TECHNICAL SUPPORT

To contact Affymetrix® Technical Support:

AFFYMETRIX, INC.
3420 Central Expressway
Santa Clara, CA 95051 USA

Tel: 1-888-362-2447 (1-888-DNA-CHIP)
Fax: 1-408-731-5441

sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd.,
Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom

UK and Others Tel: +44 (0) 1628 552550
France Tel: 0800919505
Germany Tel: 01803001334
Fax: +44 (0) 1628 552585

saleseurope@affymetrix.com
supporteurope@affymetrix.com

Affymetrix Japan K.K.
Mita NN Bldg. 16F
4-1-23 Shiba Minato-ku,
Tokyo 108-0014 Japan

Tel. 03-5730-8200
Fax: 03-5730-8201

salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com

Chapter 1

Setting Analysis Parameters

Chapter 1

Setting Analysis Parameters

This chapter describes how to set analysis parameters prior to using the Affymetrix® Tiling Analysis Software (TAS) v1.1 to analyze Affymetrix GeneChip® Tiling Array data.

Data Files

TAS uses a combination of files to perform and store analysis results, reports and graphs:

- CEL – files generated by the GCOS system
- BPMAP – genomic location files supplied by Affymetrix
- GRC – grid alignment verification files supplied by Affymetrix
- GFF – files that store positive and negative control data

For a complete description of files, see *Appendix A, TAS File Types*.

Unlike GCOS Server software, TAS can read any data file accessible on the file system. Due to the file security feature of GCOS, the CEL files managed by the GCOS software are not accessible to the TAS application. Therefore, you must export the data directly from GCOS using the Data Transfer Tool (**Start → Programs → Affymetrix → Data Transfer Tool**).

Analysis Settings

Prior to analyzing tiling data, you must specify analysis settings and report parameters in the Default Properties window. Click **Edit → Defaults** to view the Default Properties window and choose analysis settings and parameters on the following six tabbed pages:

- Data Paths
- Export
- Normalization Options
- Scale
- Probe Analysis
- Interval Analysis

NOTE **The options you select are saved from session to session and can be viewed again by reopening the Defaults Properties window.**

File Properties

File properties and settings can be viewed by selecting **File → Properties → (file name)**, and clicking **Open.** The properties are listed in the Workflow status messages pane located in the lower half of the TAS window.

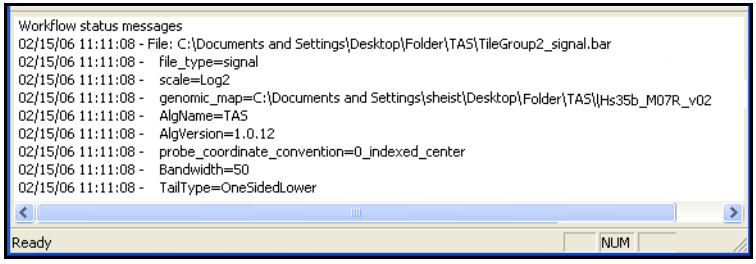

Figure 1.1
BAR file properties displayed in the lower half of the TAS window

DATA PATHS

Settings on the *Data Paths* tab (Figure 1.2) are used to specify the location of data files and BPMAP files. Data files (CEL, BAR, BED, MVA, ROC, RPT and TAG) are stored in the data directory. Library files (BPMAP and GRC) are stored in the library directory. These directories are only a starting point for locating your data files. Your TAS data and library files may be located in any directory you indicate on the file system. For description of file types, see *Appendix A, TAS File Types*.

NOTE **The most current library files can be downloaded from *www.affymetrix.com*.**

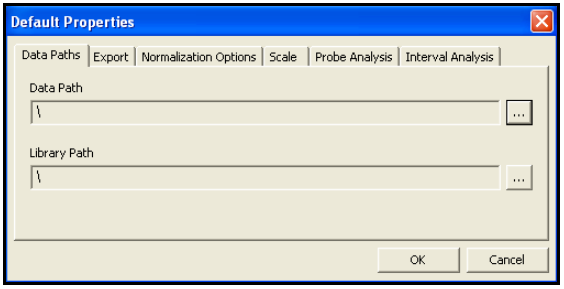

Figure 1.2
Data Paths tab

EXPORT

Settings on the *Export* tab (Figure 1.3) are used to specify the type of data saved in the BAR files and whether or not analysis results (signal and p-values) should be exported along with the BAR files as ASCII text files.

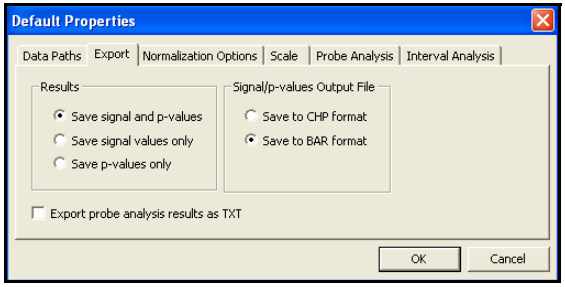

Figure 1.3
Export tab

NORMALIZATION OPTIONS

The setting displayed on the *Normalization Options* tab (Figure 1.4) is used to specify target intensity when normalizing CEL data. The intensities in the CEL file are linearly scaled so that the median intensity value is equal to the target intensity.

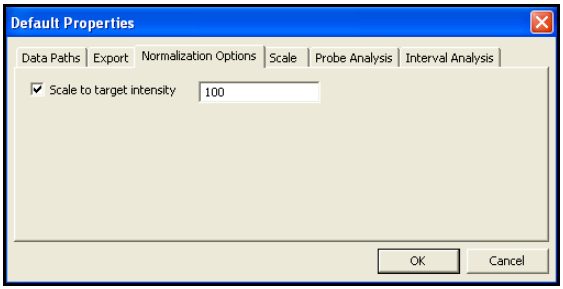

Figure 1.4
Normalization Options tab

SCALE

Settings on the *Scale* tab (Figure 1.5) are used to specify the scale in which the analysis results (signal and p-values) are to be stored in the output files.

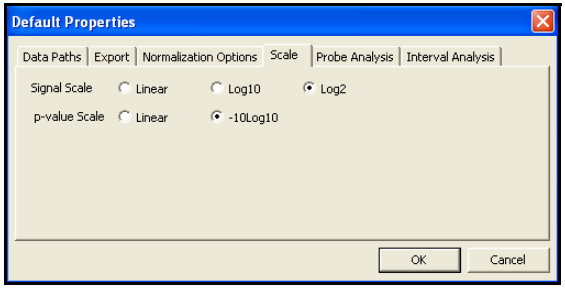

Figure 1.5
Scale tab

PROBE ANALYSIS

Settings on the *Probe Analysis* tab (Figure 1.6) are used to determine probe analysis levels for computing signal and p-values. Since bandwidth defines the number of bases to extend from the position being analyzed, these settings ensure that every probe in a region of *2*Bandwidth + 1* is included in the signal and p-value analysis.

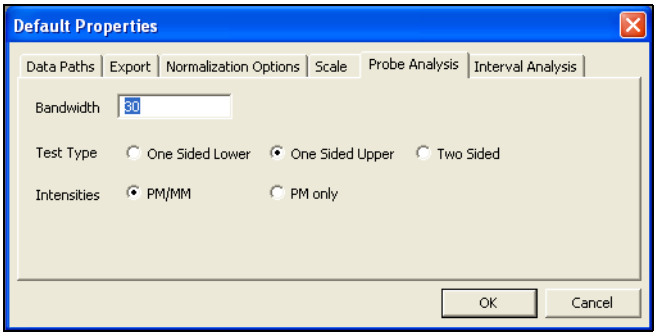

Figure 1.6

Bandwidth (BW) is determined by two main factors:

- Base pair (bp) tiling on the array
- Type of experiment

For example, in an RNA transcript mapping experiment on a 35 bp tiled array, we suggest that the BW range is 50 – 90, such that the window size (2 x BW +1) roughly corresponds to the median size of an exon (on chr. 21 & 22 = 137 bp).

Buttons on the Test Type line are described as follows:

- One Sided Lower: derives p-values to test whether or not the treatment group has a greater signal than the control group.
- One Sided Upper: derives p-values to test whether or not the treatment group has a lesser signal than the control group.

- Two Sided: derives p-values to test whether or not the treatment group has a different signal than the control group.

Buttons on the Intensities line are described as follows:

- PM/MM: Perfect Match/Mismatch
- PM only: Perfect Match

NOTE

Use the *PM/MM* button to specify that both PM and MM probe intensities are to be used in the analysis, and use the *PM Only* button when the sequence does not include a mismatch.

If the *PM/MM* button is checked and the sequence being analyzed is actually *PM only*, TAS performs a PM-only analysis for that sequence.

For more information about Test Type, see *Appendix B, Algorithms and Parameters*.

INTERVAL ANALYSIS

Settings on the *Interval Analysis* tab (Figure 1.7) are used to determine how intervals are calculated. Base pair spacing of probes on the array, type of experiment (RNA mapping or ChIP-chip), and stringency of the data set to be created determine which values to use.

In a typical RNA mapping environment, users may want to define threshold for signal interval generation based on:

- Visualization of the signal BAR file in IGB
- Varying the min/max Y-axis scale slider to generate specific signals detected in exon regions
- Minimizing excessive noise or non-specific signal in intronic or intergenic regions

For example, if the array background signal is approximately 100, the signal threshold would be set close to this number on a linear scale, or

approximately 6.64 on the log2 signal scale. Higher threshold settings would result in a lower false positive rate, but would possibly limit the detection of low abundance transcripts.

For a typical ChIP-chip experiment, for p-value interval generation, threshold is set based on the confidence level of p-values in the two-sample analysis that users select. Affymetrix has used p-value cutoffs between 10^{-3} and 10^{-5} for interval generation (30 to 50, respectively, on the -10log10, p-value scale). In both RNA mapping and ChIP-chip experiments, max gap and min run parameters are dependent on size of intervals that users want to identify.

NOTE **The scale of the threshold entered must match the scale of the data in the BAR/CHP file.**

Calculations are performed by:

- Determining the region where a probe is positive (the signal or p-value is above or below a threshold).
- Selecting a maximum gap between positive probes.
- Selecting a minimum length or run of adjacent probes.

BAR files and corresponding BED files contain the computed intervals. The Interval Analysis function uses data stored in a signal or p-value BAR file. For instructions on performing Interval Analysis, see *Interval Analysis, on page 23*.

For more information about Threshold, Maximum Gap, and Minimum Run, see *Appendix B, Algorithms and Parameters*.

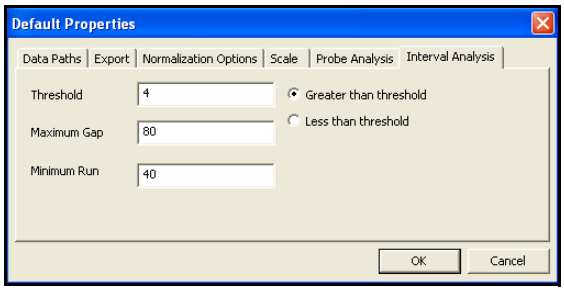

Figure 1.7
Interval Analysis tab in the Default Properties window

Analysis Group

After you have properly configured TAS for your experiment, you are ready to define the data files for analysis. See *Chapter 2, Defining The Analysis Group*.

Chapter 2

Defining The Analysis Group

Chapter 2

Defining The Analysis Group

This chapter describes step-by-step instructions on how to define the group of files to be analyzed with the Affymetrix® Tiling Analysis Software (TAS).

Analysis Group

The analysis group is the set of CEL files you select for either a one sample detection analysis or a two sample comparison analysis and is analyzed to produce signal and p-values for each genomic probe position.

SELECT THE BPMAP

To define the BPMAP:

1. Select **File → New.**
   The TileGroup window opens.

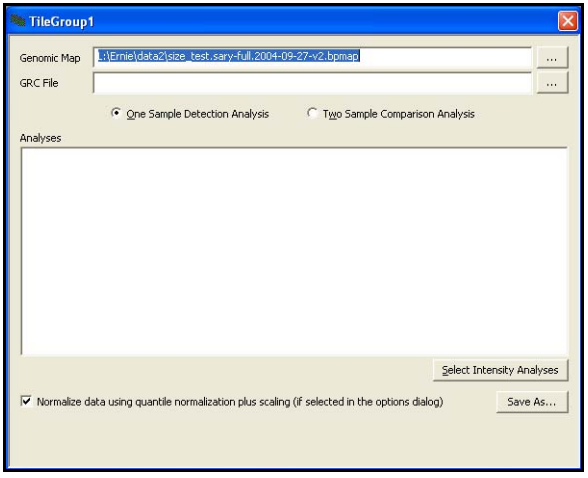

Figure 2.1
Genomic Map for the location of the BPMAP file

2. Use the browse button [...] to select the **Genomic Map**. This is the BPMAP file supplied with the tiling array used for your experiment.
3. Use the associated browse button to select the **GRC** file for the reporting function.

NOTE **If a GRC file is not available, leave the field blank. The GRC file is used only in the Report function and is optional.**

4. Select either the **One Sample Detection Analysis** or **Two Sample Comparison Analysis** radio button and follow the steps below for each selection.

ONE SAMPLE DETECTION

1. Select the **One Sample Detection Analysis** radio button.

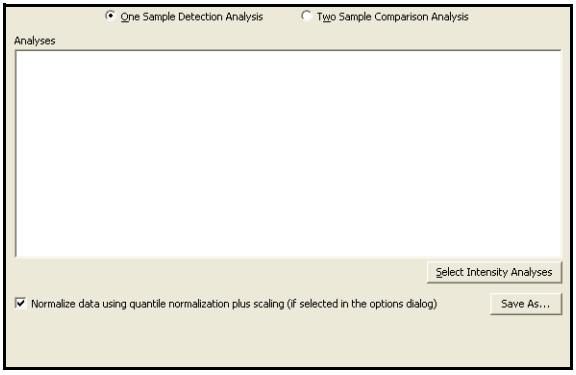

Figure 2.2
One Sample Detection Analysis

2. Select the normalization checkbox, if appropriate. When you check this option, TAS includes the normalization process as part of the analysis.

NOTE

You can delete individual CEL files from the group by selecting the file in the list and pressing the *Delete* key.

3. Click the **Select Intensity Analyses** button at the bottom of the screen, and browse for your CEL files.

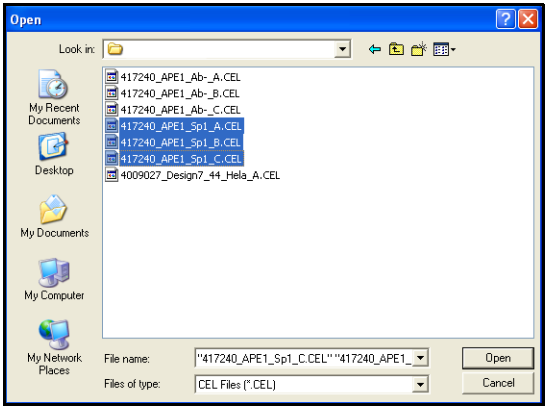

Figure 2.3
Group of CEL files for a one sample detection analysis

4. Click **Open**. The selected group appears in the analysis window.

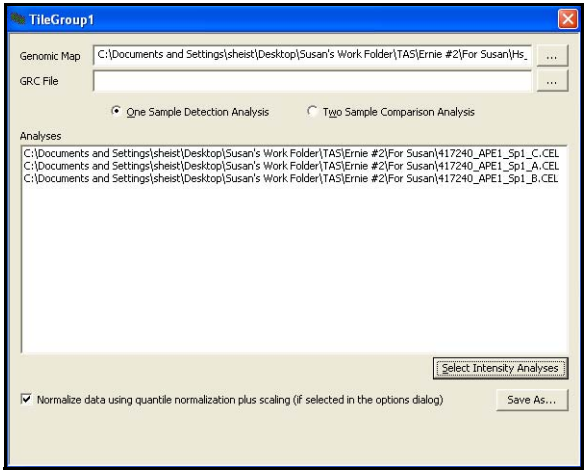

Figure 2.4
One sample analysis group of selected CEL files

5. Click **Save As** to save the group as a TAG file in your specified directory.

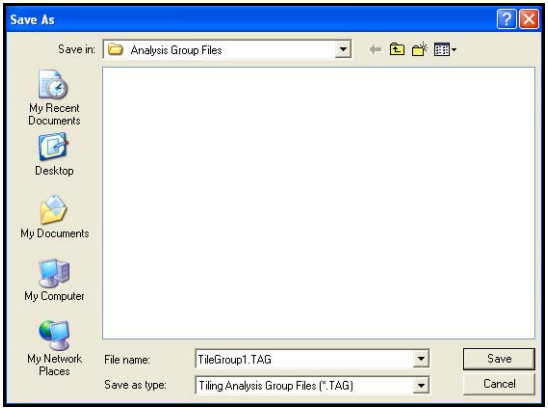

Figure 2.5
Save as TAG file

6. After you have selected analysis parameters (*Chapter 1, Setting Analysis Parameters*) and defined the analysis group, proceed to *Chapter 3, Analyzing Intensity Data*.

TWO SAMPLE COMPARISON

1. Click the **Select Treatment Group Analyses** button and select the CEL files from the treatment group (Figure 2.6).
2. Click the **Select Control Group Analyses** button and select the CEL files from the control group (Figure 2.6).

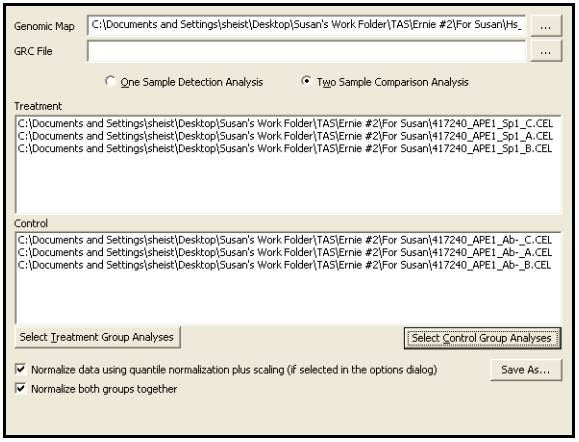

Figure 2.6
Treatment Group and Control Group for two sample comparison

3. Select the normalization checkboxes, if appropriate. When you check these options, TAS includes the normalization process as part of the analysis.

NOTE **You can delete individual CEL files from the group by selecting the file in the list and pressing the *Delete* key.**

4. Click the **Save As** button and save the group of files to a TAG file.

Analyzing Intensity Data

After you have selected the analysis parameters (*Chapter 1, Setting Analysis Parameters*) and defined the analysis group, proceed to *Chapter 3, Analyzing Intensity Data*.

Chapter 3

Analyzing Intensity Data

Chapter 3

Analyzing Intensity Data

This chapter gives step-by-step instructions on analyzing intensity data with the Affymetrix® Tiling Analysis Software (TAS). The following sections describe the various analyses that you can perform with TAS.

- Intensities
- Intervals
- Promoter
- Enrichment
- Report
- Interval Overlap Report
- ROC Plot
- MVA Plot

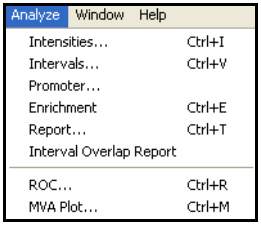

Figure 3.1
Analysis menu

Probe Analysis

Probe analysis (**Analyze → Intensities**) uses a combination of CEL files, the BPMAP file, normalization settings specified in the TAG file, and parameters specified in the Default Properties window. Probe analysis results in a signal and a p-value for each genomic position interrogated by the array. These values are then stored in two BAR/CHP files (Figure 3.2).

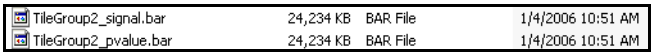

Figure 3.2
Example of signal and p-value BAR files

To perform intensity analysis:

1. Click **Edit → Defaults** and verify that the correct parameters are set in the Default Properties window. Refer to *Chapter 1, Probe Analysis (see page 7)*.
2. Click **Analyze → Intensities** (or click).
   The **Select TAG files to analyze** window opens.

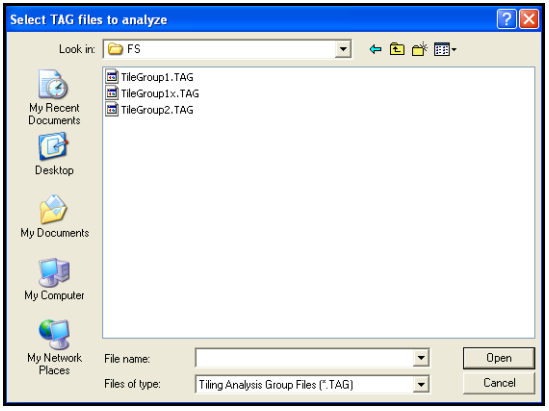

Figure 3.3
Select TAG files to analyze window

3. Select the appropriate TAG file and click **Open** to start analysis.

NOTE

During data analysis, status messages display in the *Status message window* in the lower half of the main TAS window. When the process is complete, completion date and time are displayed.

Interval Analysis

Interval analysis computes the intervals from data stored in the BAR (signal or p-value) or CHP file.

To perform interval analysis:

1. Click **Edit** → **Defaults** and verify that the interval analysis parameters are set in the Default Properties window. Refer to *Chapter 1, Interval Analysis (see page 8)*.
2. Click **Analyze** → **Intervals** (or click ). The **Select BAR/CHP files to compute intervals** window opens.
3. Select the appropriate BAR or CHP file to be analyzed and click **Open** to start the analysis.

NOTE

The scale of the threshold entered on the Interval Analysis tab must match the scale of the data in the BAR file. Refer to *Real-Time Intervals (see page 25)* to optimize the interval analysis settings.

The BAR file creates a corresponding BED file that contains the computed interval (Figure 3.4). These are displayed in the **Workflow status messages** pane during analysis.

NOTE

TAS, versions 1.0.14 and higher, changed the method for defining the start and stop positions. Each position is now relative to the center base of the probe. For a detailed description of this updated convention, go to the following web address:

http://www.affymetrix.com/support/developer/downloads/TilingArrayTools/probe_coordinate_convention.affx

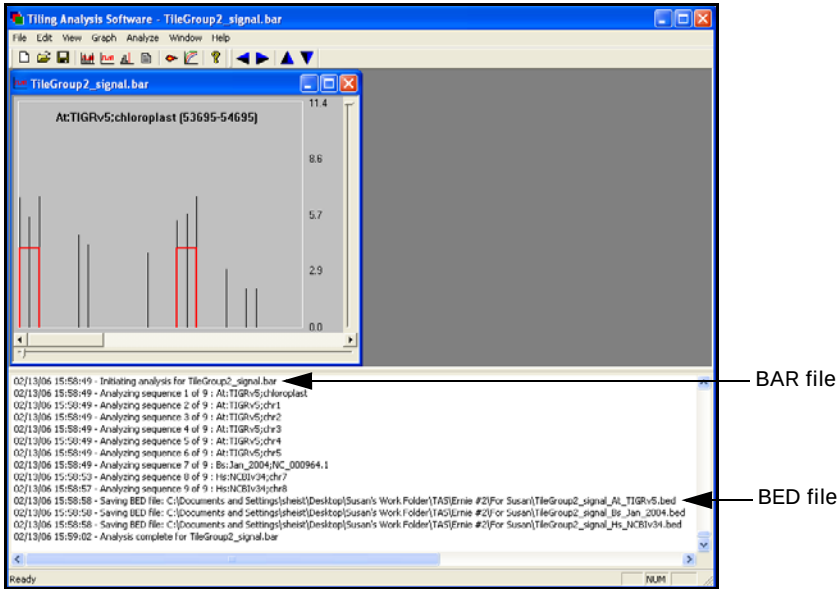

Figure 3.4
Interval analysis performed on a BAR file creates BED files

4. Select **File → Properties**.

The **Select files to view properties** window opens.

5. Select the BAR file and click **Open** to view scale and other settings (Figure 3.5).

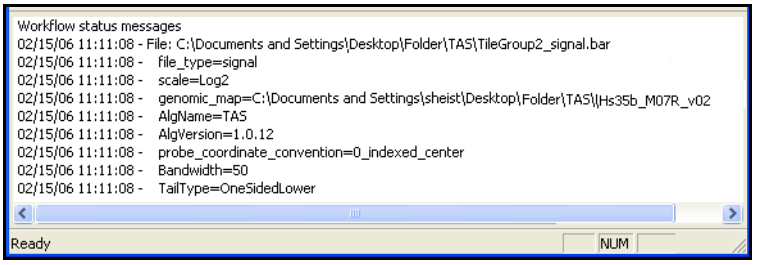

Figure 3.5
BAR file properties displayed in the lower half of the TAS window

Real-Time Intervals

To display real-time intervals:

1. To see a preview of the intervals computed in real time as the parameters are modified, open the **BAR file viewer** :

A. Click **File** → **Open** and select the **BAR** file.

The signal or p-values stored in the BAR file are displayed in a bar graph one sequence at a time.

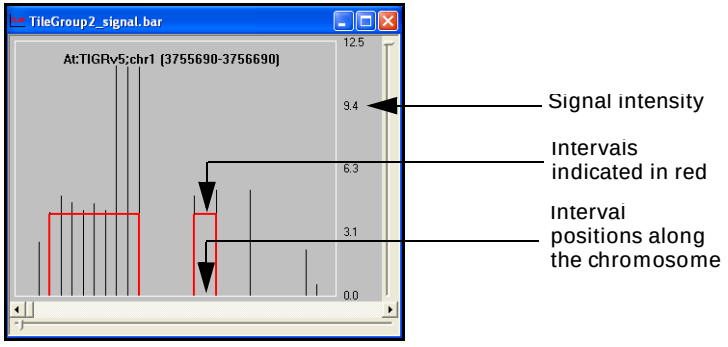

Figure 3.6
BAR file graph

B. Use the Graph menu command or the up and down arrow keys in the toolbar to view subsequent sequence results. The slider adjusts the amount of data to be displayed.

Figure 3.7
Graph menu and toolbar arrows

2. To adjust the BAR file in real time, click **Graph → Adjust Defaults.**

The Interval Parameters dialog box opens.

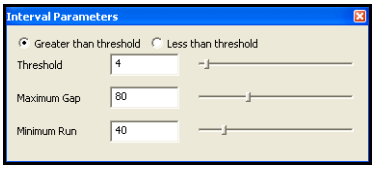

Figure 3.8
Interval Parameters dialog box for real time viewing

When you change the parameter settings in this dialog box, the BAR graph updates in real time to display the computed intervals.

NOTE

This capability does not affect the parameters used when creating BED files. To change BED file output, you must specify interval parameters on the Interval Analysis tab in the Default Properties window (*Edit → Defaults*).

Intervals are only computed for the visible portion of the BAR data.

Promoter Analysis

Promoter analysis computes data stored in a BED file and a reference promoter BED file. Promoter Analysis is designed to join detected promoter regions from an experiment Interval BED file, to their associated downstream gene regions in a Promoter BED file that contains the probe set IDs representing genes on the U133 expression arrays. Detected regions that are associated with the U133 probe sets are output to the Results BED file. Using this analysis tool, researchers can correlate detected regions from tiling array applications (such as chromatin immunoprecipitation studies) with their respective gene expression level.

To perform promoter analysis:

1. Click **Analyze → Promoter**.

   The Promoter Analysis window opens.

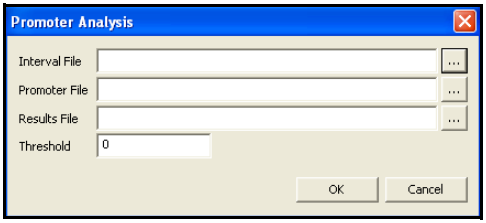

Figure 3.9
Promoter Analysis window

2. Click the corresponding [...] buttons to select the Interval BED file, the Promoter BED file, and the Results BED file.
   - Interval BED: The file generated by the experiment.
   - Promoter BED: A library file produce by a database query.
   - Results BED: Regions detected in an output BED file used to select regions of the library file and stored in the Results BED file. This file has the same format as the library file.
3. Enter a number for the threshold value.

Threshold is a number between 0 and 1000 (default is 0). As explained below, the fraction of overlap between each region in the reference promoter and the detected regions is calculated and scored as a number between 0 (no overlap) and 1000 (100% overlap). Regions with scores lower than the threshold are not reported.

NOTE

The threshold value entered is compared to the overlap score. If an interval has an overlap score greater than the threshold, then the promoter reference interval with computed overlap score is added to the output BED file.

The resulting intervals, associated probe set identifiers, and overlap scores are stored in a BED file.

4. Use the following algorithm for each interval specified in the promoter reference BED file:

   A. Compute the sum of the overlaps between the promoter reference interval and all the intervals in the input BED file.

   B. Verify that the overlap score is 1000* overlap/ promoter interval size. This score is also used by the genome browser to emphasize regions in the display – the higher the score, the greater the emphasis.

NOTE

If the overlap score is greater than the input threshold, then the promoter reference interval (with computed overlap score) is stored in the output BED file.

Enrichment Analysis

Enrichment Analysis performs a simulation to assess the likelihood that the observed overlap is different from the expected overlap with no enrichment. The result is a histogram where no enrichment is the assumed case (random intervals). The histogram contains a vertical line marking the observed value (the enrichment).

To perform enrichment analysis:

1. Click **Analyze** → **Enrichment** (or click ).
   The Enrichment Analysis window opens.

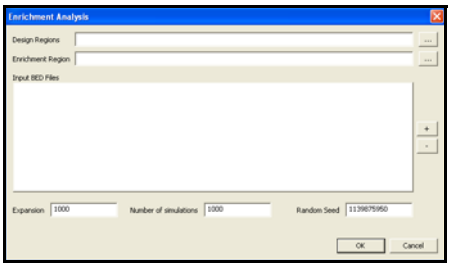

Figure 3.10
Enrichment Analysis window

2. Enrichment regions are specified in the BED file.
3. Click the  button and select the appropriate BED file(s) to analyze. Use the  button to remove BED files from the list.
4. In the **Expansion** field, specify the number of bases to expand the design regions.
5. In the **Number of simulations** field, specify the number of simulations to run.
6. In the **Random Seed** field, enter the random number generator.

Two histograms, one for baseline and one for interval, are displayed for each input BED file you entered. The histograms contain a vertical marker to mark the enrichment.

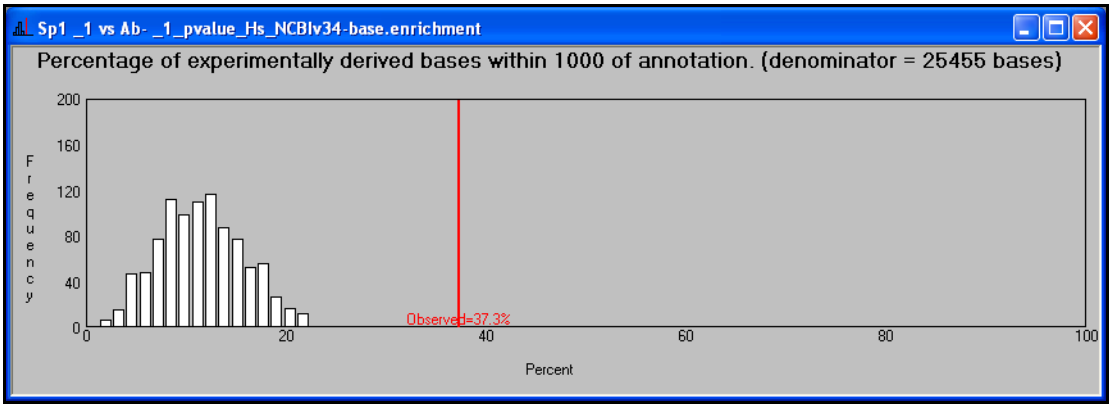

Figure 3.11
Example of a base enrichment

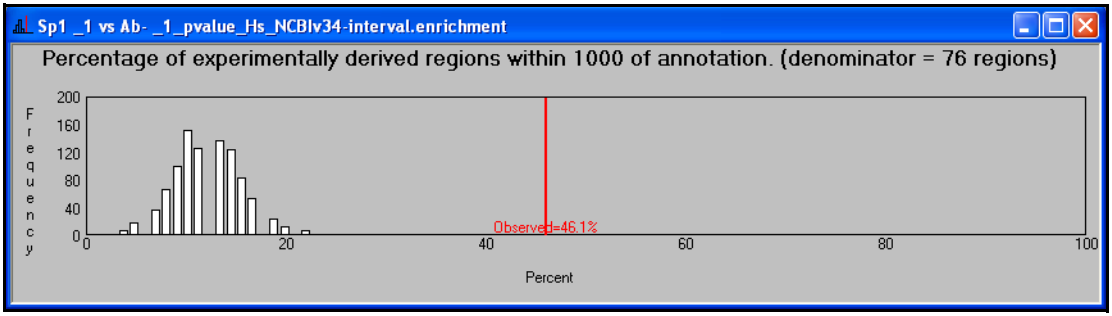

Figure 3.12
Example of an interval enrichment

Report Files

The report functionality provides an assessment of intensity data stored in a CEL file.

To create a Report file:

1. Click **Analyze** → **Report**.
2. Select a TAG file and click **Open**.
3. When prompted to turn off the normalization specified in the TAG file, click **Yes**.

NOTE **This provides the ability to run the report with or without normalization settings without having to change the TAG file.**

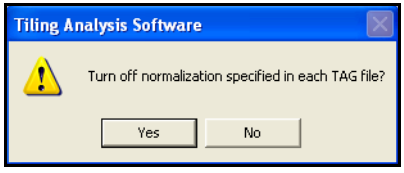

Figure 3.13
.RPT file prompt to turn off normalization in TAG file

NOTE **All CEL files stored in the TAG file are automatically analyzed and stored in a RPT file.**

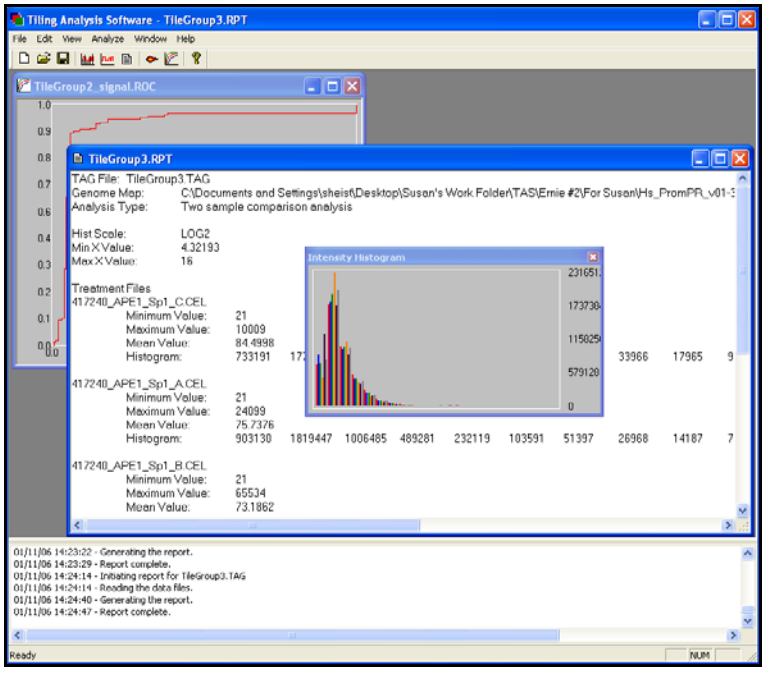

Figure 3.14
Report File (RPT)

For each CEL file, the report includes:

- Minimum intensity value
- Maximum intensity value
- Mean intensity value
- The number and percentage of bright controls that failed the intensity test (if a GRC file was defined in the TAG file)
- The number and percentage of dim controls that failed the intensity test (if a GRC file was defined in the TAG file)
- A histogram of the log base 2 of the intensities

  There are 20 bins in the histogram.

Interval Overlap Report

The Interval Overlap Report provides the number of intervals and bases from a BED file that are within a user-specified distance of intervals in a reference BED file.

In an analysis where the user specifies the distance as 0, the report displays the number of intervals in the BED file that overlap any interval in the reference BED file.

To display the Interval Overlap Report:

1. Select **Analyze → Interval Overlap Report.**
   The **Select the test BED file** window opens.

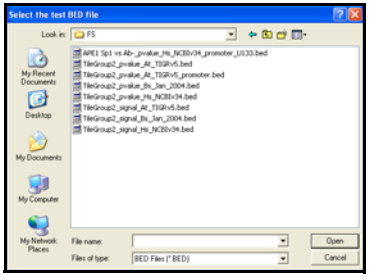

Figure 3.15
Select the test BED file

2. Select the test BED file and click **Open**.
   The **Select the baseline BED file** window opens.

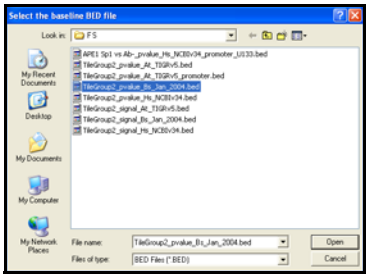

Figure 3.16
Select the baseline BED file

3. Select the baseline BED file and click **Open**.
4. Enter the number of bases to expand the baseline intervals in the popup window and click **OK**. This refers to the number of bases added to each side of the intervals in the reference BED file.

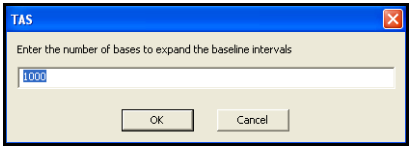

Figure 3.17
Enter the number of bases

An Interval Overlap Report is generated.

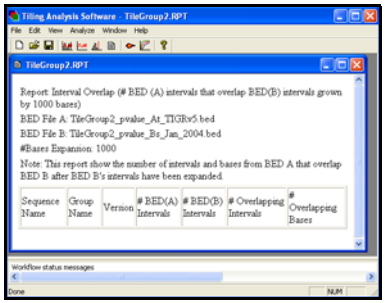

Figure 3.18
Interval overlap report

IMPORTANT !

The order of the input BED files is important. BED1 versus BED2 yields different results than BED2 versus BED1.

The Interval Overlap report includes:

- The input BED files and expansion size
- The number of intervals in the test BED file
- The number of intervals in the reference BED file
- The number of overlapping intervals
- The number of overlapping bases

ROC Plot

The ROC (receiver operator characteristics) is used to calculate the false positive rate and sensitivity of a given experiment for a specific threshold and the signals or p-values stored in a BAR file. The list of positive and negative controls for the ROC analysis is obtained from the information in the GFF file.

To generate the ROC curve:

1. Click **Analyze** → **ROC** to open the ROC Options dialog box.

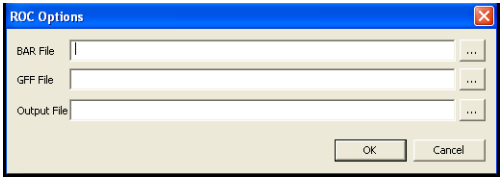

Figure 3.19
ROC Options dialog box

2. When prompted, select the **BAR** file, **GFF** file and output file name.

   The results are stored in a ROC file (text formatted file) and displayed in the application.

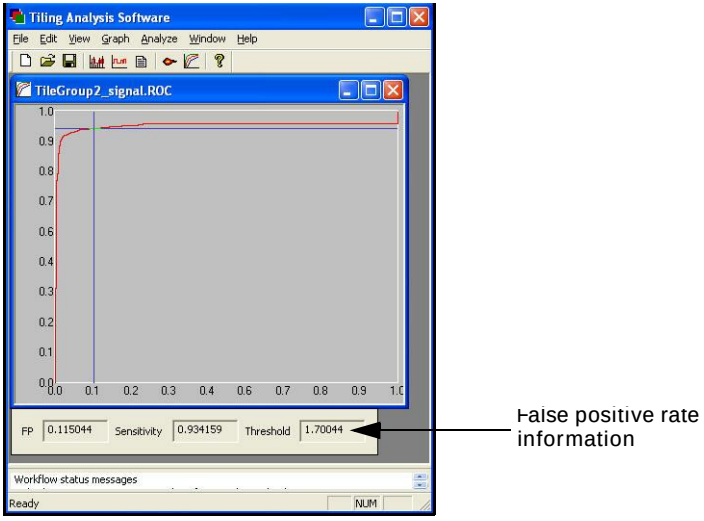

Figure 3.20
ROC plot

3. Select **Graph → False Positive Rate** (or click and drag across the window and click again) to view the false positive rate, sensitivity, and threshold.

MVA Plot

The (MVA) Plot is an X/Y scatter plot that uses the intensities from two CEL files. The X axis is defined as the average of the log base 2 of the intensities from the two CEL files, and the Y axis is the difference of the log base 2 of the intensities of the two CEL files.

The BPMAP file is optional. If omitted, then all the data in the CEL files are plotted. If provided, then only those intensity values that correspond to a tiling sequence item will be used.

To generate an MVA plot:

1. Click **Analyze → MVA Plot** to generate an MVA plot.
2. When prompted, select two or more CEL files, the display threshold and the normalization option.

   A graph is displayed for every pair combination of CEL files, and each of the CEL file-pair results are stored in a named MVA file. The MVA file may be opened at a later time to view the plot.

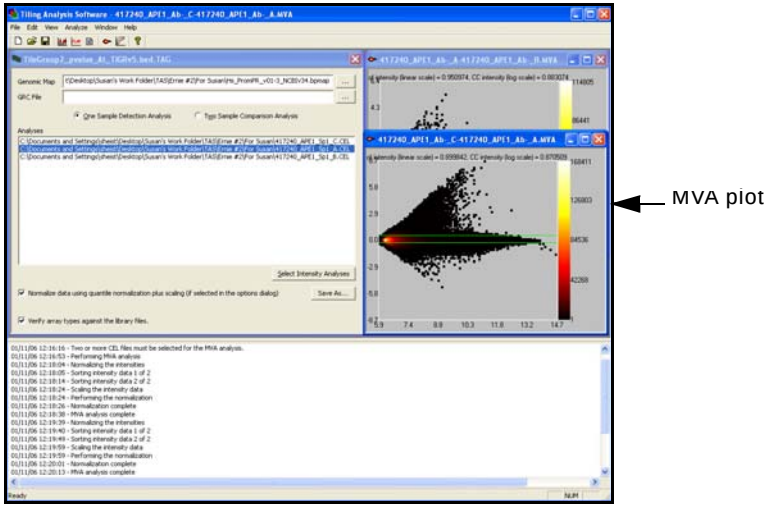

Figure 3.21
MVA plot analysis

The MVA Plot is displayed in one of three different color scales:

- Black bar
- Grey
- Blue/red

Use the context sensitive menu (right click in the .MVA window) to select a different color scale.

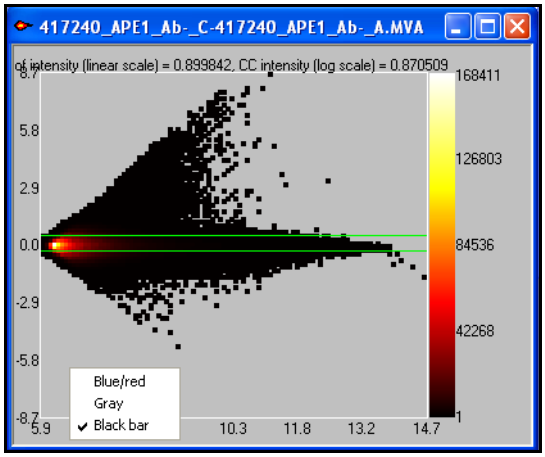

Figure 3.22
Right click for plot color choices

Modifying the Track in the BED File

The default track in a BED file is *AFFX*. You can modify this value by doing the following:

1. Click **File** → **Modify Track**.
2. Enter the new track in the **Track** field.
3. Click the Add button to specify the BED files to modify.
4. Click OK to modify the specified BED files with the new track name.

NOTE

Use the *REMOVE* button to remove any selected BED files from the list.

Appendix A

TAS File Types

Appendix A

TAS File Types

The Affymetrix® Tiling Analysis Software uses or generates the following files:

File Type Definitions

CEL FILES

The CEL (Cell Intensity File) file contains processed cell intensities from the primary image in the .DAT file. Intensity data is computed by the Affymetrix® GCOS application and stored in CEL files. Each GeneChip® Tiling Array produces a single CEL file.

BPMAP FILES

BPMAP (Binary Probe Map) files contain the genomic probe position map. The BPMAP file maps the X/Y coordinate of a probe on a GeneChip array to a genomic position for an intended function. It designates a probe as either a perfect match (PM) or a mismatch (MM) probe. The mapping between probe and genomic position may change as the genome annotations are revised.

TAG FILES

A TAG (Tile Analysis Group) file stores a set of CEL files, a BPMAP file, and the normalization option settings for a single one or two sample analysis.

ROC FILES

The ROC (Receiver Operator Characteristics) file stores the false positive rate, sensitivity measurements, and the thresholds used to draw an ROC curve. The ROC file is the result of ROC analysis, and the format is a simple, tab-delimited text file.

MVA FILES

The MVA file stores data needed to generate an MVA plot.

BAR/CHP FILES

The BAR (Binary Analysis Results) and CHP files store the probe signal and p-values calculated by the software. These file types may also be used to store CEL file intensities organized by the structure defined in a BPMAP file.

BED FILES

The BED file stores the start and stop positions of each computed interval. Visit *http://www.genome.ucsc.edu/FAQ/FAQformat* for more information on the BED file format.

RPT FILES

The RPT (Report) file stores report results. This is displayed as an ASCII text file.

GFF FILES

The GFF (General Feature Format) file is an exchange format for feature descriptions. The GFF file is used to store the set of positive and negative controls for an ROC analysis. Visit *http://www.sanger.ac.uk/software/formats/GFF/* or *http://www.genome.ucsc.edu/FAQ/FAQformat* for more information on the GFF file format.

NOTE 

File format documentation and software development kits to parse data files are available at:

http://www.affymetrix.com/genechip/developer

Appendix B

Algorithms and Parameters

Appendix B

Algorithms and Parameters

Algorithms

NORMALIZATION

The default workflow assumes that the probe intensity data are normalized using quantile normalization [1]. Quantile normalization makes the assumption that the data being normalized have the same underlying distribution; this should be a reasonable assumption within biological and sample replicates. For the one-sample analysis, the assumption of equal, underlying distributions is usually reasonable. By default, all arrays are quantile-normalized together. For two-sample analysis, it is quite possible that the underlying distributions are different for the two groups; therefore, by default, quantile normalization is performed only within each group.

Normalization, if performed, is full quantile normalization of all probe intensities. The user can select the option to normalize TAG files separately for each group or normalize across all groups in an analysis. If the target intensity parameter is set, then the normalized intensities are further scaled to set the median intensity for every array to be the target intensity value. If normalization is not performed, the target intensity parameter is ignored.

PROBE ANALYSIS

The BPMAP file is used to associate each perfect match (PM) probe with its position in a genomic sequence. For example, in the GeneChip® Human Promoter 1.0R Array, a PM probe is typically associated with a chromosome. The probe position, whether its target is the forward or the reverse strand, is determined by the location of its 0-based position on the lower coordinate of the probe aligned to the target. Probe position is defined by the positions employed in the genome assembly, which is used for probe selection in the array; for example, the GeneChip® Human Promoter 1.0R Array contains probes from assembly NCBIv34. The version information is reported in the viewer of output BAR files and in comments embedded in output BED files. Mismatched (MM) probes are always paired with a PM probe and have the same convention.

Once the PM and MM pairs have been associated with sequence positions and normalized, the next step is to perform statistical analysis (in a local context for each position) to determine size and significance of the hybridization signal. In the case of **PM only** analysis, MM is not used. On some arrays, MM is not present.

Analysis can be performed in either a one-sample or a two-sample context. A typical one-sample context might consist of using a number of biological or technical replicates for detection of regions of transcription. A typical two-sample context might consist of a treatment versus control comparison to look for regions of enrichment in a chromatin immunoprecipitation (ChIP) experiment.

Analysis (known in TAS as *Probe Analysis*) is focused at a single sequence position because the method is the same for all positions. The first step is to define a local data set consisting of all PM probes located within ± bandwidth base pairs of the position of interest. The value of the bandwidth should be driven by the average size in base pairs of the signal to be detected. In the case of transcription monitoring, the bandwidth would typically be on the order of half the average exon length, often about 50bp. In the case of ChIP assays, it would be half the expected fragment length in the step immediately before enrichment, which is assay dependent, but typically on the order of 500bp.

Selection of a bandwidth involves a tradeoff. On one hand, the bandwidth should be as large as possible to provide greatest statistical power for the analysis at each position; on the other hand, if the bandwidth is too large, the analysis tends to dilute signal by including background. The resulting local data set typically consists of a number of PM probes for each array being studied. The next step differs for one-sample and two-sample analysis.

ONE-SAMPLE ANALYSIS

In a one-sample analysis, TAS performs a Wilcoxon signed-rank test on the *n* probe intensity differences {PMj-MMj; i=1,...n} by testing the null hypothesis of no shift between the distribution of PM intensities and MM intensities. The default alternative hypothesis is that there is a positive shift in the distribution of PM-MM, and therefore, a one-sided p-value is reported for the position. The type of p-value reported can be changed in TAS on the Scale tabbed page (**Edit → Defaults → Scale**). The p-value reported in the output file may be $-10log_{10}(p\text{-}value)$, which is a more suitable quantity for plotting against sequence position; higher values are more significant. This converts a p-value of 0.1 to a transformed p-value of 10, 0.01 to 20, 0.001 to 30, and so on (this is the same transform as the one used for Phred quality scores in the DNA sequencing literature).

An estimate of signal intensity is also computed. The estimator used is the Hodges-Lehmann estimator [2] which is the usual estimator associated with the Wilcoxon signed-rank test, and which is also known as the *pseudomedian*. After forming all *n* values $\{D_i = PMj\text{-}MMj; i=1,...,n\}$, the $n(n+1)/2$ pairwise averages $(D_i\text{-}D_j)/2$, known as Walsh averages, are computed. The estimate *s* of signal location is taken to be the median of the $n(n+1)/2$ Walsh averages and is then transformed to $\log_2(\max(s,1))$.

The one-sample analysis tests whether or not the PM probes have significantly greater signal than their corresponding MM probes and therefore is used only if MM probes are present. In arrays where PM-only probes are present, the analysis will run, but the result is not useful.

TWO-SAMPLE ANALYSIS

In two-sample analysis, there are two data sets, which are called a treatment and a control group. Each group consists of the subset of data falling within the specified bandwidth as described above, resulting in n_t treatment pairs of probe intensities $\{PM_{t,i}\text{-}MM_{t,i}; i=1,...n_t\}$ and n_c control pairs of probe intensities $\{PM_{c,i}\text{-}MM_{c,i}; i=1,...n_c\}$. The log-transformed quantities $\{S_{g,i}=\log_2(\max(PM_{g,i}\text{-}MM_{g,l},1)); g=t,c;i=1,...,n_g\}$ are formed and a Wilcoxon signed-rank test is performed on the two samples $\{S_{t,i};i=1,...,n_t\}$ and $\{t_{c,i};i=1,...,n_c\}$. In the case of a **PM only** analysis, instead of using the log-transformed differences, the log-transformed PM signal intensities $\{S_{g,i}=\log_2(PM_{g,i});g=t,c;i=1,...,n_g\}$ are used.

The default test type is a one-sided test, against the alternative that the distribution of the treatment data is shifted up with respect to the distribution of the control data. A two-sided or lower-sided test can be used instead of the one-sided lower **(Edit → Defaults → Probe Analysis)**. Similar to the one-sample p-values, by default, the $-10\log_{10}$ transform is applied to the output to enable visualization along the sequence.

An estimate of fold enrichment is also computed; the estimator used is the Hodges-Lehmann estimator associated with the Wilcoxon rank-sum test [2]. The estimator is computed by forming all n_tn_c values $\{D_{ij}=(S_{t,i}\text{-}S_{c,j});i=1,...,n_t;j=1,...,n_c\}$. The Hodges-Lehmann estimator is then the median of the D_{ij} and can be interpreted as the $\log_2$ fold change between the treatment and control group signals.

INTERVAL ANALYSIS

In both the one-sample and two-sample analysis, the *Probe Analysis* step described above will yield a p-value and a signal estimate associated with the location of each position in the sequence to which a probe pair aligns. TAS writes the resultant signals to output files, which can then be viewed in the Integrated Genome Browser (IGB). Additionally, these signals can be thresholded to produce discrete regions, which meet certain detection criteria, along the sequence of interest.

The method involves three steps:

- In the first step, a threshold is applied to the value at each probe position, and a position is classified as positive if its value exceeds the threshold. The threshold can be applied to the signal, and a position can be classified as positive if it is either greater than or less than the threshold supplied. Alternatively, the threshold can be applied to the p-value associated with each position, in which case, one is typically interested in positions with p-values lower than the threshold. Use the Interval Analysis tabbed page (**Edit → Defaults → Interval Analysis**) to set the threshold for p-value or signal and to look for positions larger or smaller than the threshold.
- In the second step, positive positions are separated by a distance of up to *max_gap* are joined together to form detected regions. The choice of *max_gap* is up to the user and depends on assay conditions. In general, making it larger is more permissive and will be more forgiving of positions which failed to make the threshold in a run of otherwise positive positions.
- The final step is to process the list of all detected regions and reject any with a length of less than *min_run*. Again, the choice is dependent on the assay used, but generally making *min_run* smaller is more permissive and allows for shorter runs of positive positions to be classified as detected. The final set of all detection regions is written to an output file and can be used as a starting point for downstream analysis.

Adjustable Parameters

The following parameters are adjusted by the user on the **Probe Analysis** tabbed page (**Edit → Defaults → Probe Analysis**):

BANDWIDTH

Bandwidth is a distance (in base pairs) used to locally group positional data. Analysis at a particular position will be based on all data aligning within ± bandwidth of the position, so that the sliding window of the analysis is of size *2*bandwidth+1*.

Making bandwidth larger brings more data into each test providing more statistical power and a greater ability to detect signal. However, once the bandwidth exceeds the point where the window is larger than the signal being interrogated, power decreases to include data with no signal.

TEST TYPE

This type of test can be one-sided lower, one-sided upper, or two-sided depending on the kind of signal the user wants to locate. In one-sample analysis, use only the one-sided upper, since it is not meaningful to do large scale testing against the alternative hypothesis that the MM signals are larger than the PM signals.

INTENSITIES

If **PM/MM** is checked, then PM/MM probe pairs will be used (if present on the array). If **PM only** is checked, then PM-only intensities will be used.

The following parameters are adjusted on the **Interval Analysis** tabbed page (**Edit → Defaults → Interval Analysis**):

THRESHOLD

This parameter is applied to either estimated signal or p-value by selecting the corresponding radio button. The threshold is applied to detect positions with values either greater than or less than the supplied threshold via the **Greater than threshold** or **Less than threshold** radio buttons. For p-values, it typically makes sense to detect positions with p-values less than the threshold.

MAXIMUM GAP

This setting is the maximum tolerated gap (in base pairs) between positive positions in the derivation of detected regions. Decreasing *max_gap* results in a more stringent map; increasing it lessens the stringency.

MINIMUM RUN

This setting is the minimum size (in base pairs) of a detected region. Increasing *min_run* disallows detection of smaller regions which may be appropriate if the expected size of detected regions is large.

References

1. Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193 (2003).

2. Hollander, M. and Wolfe, D.A. *Nonparametric Statistical Methods*, 2nd edition. John Wiley and Sons, Inc. (1999).

Index

Index