



User's Guide

Affymetrix[®] miRNA QCTool

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

Trademarks

Affymetrix®, Axiom™, MyDesign™, Command Console®, DMET™, GeneAtlas™, GeneChip®, GeneChip-compatible™, GeneTitan®, Genotyping Console™, NetAffx®, and Powered by Affymetrix™ are trademarks or registered trademarks of Affymetrix Inc. All other trademarks are the property of their respective owners.

Limited License Notice

Limited License. Subject to the Affymetrix terms and conditions that govern your use of Affymetrix products, Affymetrix grants you a non-exclusive, non-transferable, non-sublicensable license to use this Affymetrix product only in accordance with the manual and written instructions provided by Affymetrix. You understand and agree that except as expressly set forth in the Affymetrix terms and conditions, that no right or license to any patent or other intellectual property owned or licensable by Affymetrix is conveyed or implied by this Affymetrix product. In particular, no right or license is conveyed or implied to use this Affymetrix product in combination with a product not provided, licensed or specifically recommended by Affymetrix for such use.

Patents

Software may be covered by one or more of the following patents: U.S. Patent Nos. 5,733,729; 5,795,716; 5,974,164; 6,066,454; 6,090,555; 6,185,561; 6,188,783; 6,223,127; 6,228,593; 6,229,911; 6,242,180; 6,308,170; 6,361,937; 6,420,108; 6,484,183; 6,505,125; 6,510,391; 6,532,462; 6,546,340; 6,687,692; 6,607,887; 7,062,092; 7,451,047; 7,634,363; 7,674,587 and other U.S. or foreign patents.

Copyright

© 2011 Affymetrix, Inc. All Rights Reserved.

Contents

Chapter 1	Introduction	3
	Conventions Used in This Guide	3
	Resources	5
	Documentation	5
	Technical Support	5
Chapter 2	miRNA QCTool Software Installation and First Steps	7
	Installing the miRNA QCTool Software	7
	Software Requirements	7
	Minimum Hardware Recommendations	7
	Software Installation	7
	First Steps	9
	Starting the miRNA QCTool Software	9
	Software Settings	9
	Window Arrangement	10
	Selection of Input Data: CEL Files	11
Chapter 3	Menus	13
	File Menu	13
	Loading CEL Files	13
	Tables Menu	13
	Data Tables / Intensities	14
	Project Description Table	15
	Quality Control Table	15
	Pearson Correlation Coefficient Table	16
	Graphs Menu	16
	Box Plots	17
	Histogram Plots	18
	MvA Plots	19
	Quality Control Plots	20
Chapter 4	Analysis	21
	Processing Steps	21
	Running an Analysis	22
	Workflow Editing	23

Appendix A	Algorithms	25
	Detection	25
	Detection Algorithm Description	25
	Wilcoxon Test	25
	Affymetrix Test	25
	Background Adjust	26
	Algorithm Description	26
	Wilcoxon Rank-Sum Test for Detection	26
	Algorithm Description	26
<hr/>		
Appendix B	Additional Information	27
	Q&A	27
	Definitions	27

Introduction

The Affymetrix® miRNA QCTool enables a variety of analysis methods tailored to miRNA studies and features extensive quality control tools. Several methods of probe signal background subtraction, array normalization (including the use of internal normalization controls), and probe set summarization are allowed, in various combinations. If you are using internal normalization controls, you can apply the default probe sets (human 5.8S rRNA) or choose from any set of the probe sets on the array, including snoRNAs, miRNAs (any species), or spike-in controls. The miRNA QCTool is thus applicable to a wide variety of experimental designs.

There are 2 versions of the tools that are available on the Affymetrix website. Please choose the correct version depending on array type (see [Table 1.1](#)).

Table 1.1 miRNA QCTool version for Array Types

Affymetrix® Array Type	Tool to Use
miRNA v1.0 Array	miRNA QCTool, version 1.0.33.0
miRNA v2.0 Array	miRNA QCTool, version 1.1.1.0 or greater

! **IMPORTANT:** Data from the GeneChip® miRNA 2.0 Arrays use will need to install miRNA QCTool version 1.1.1.0. If the previous version of the miRNA QCTool (version 1.0.33.0) has been installed on your computer, please un-install version 1.0.33.0 and reboot the computer before installing version 1.1.1.0. This is required because GeneChip miRNA 2.0 Arrays are not compatible with miRNA QCTool version 1.0.33.0 and GeneChip® miRNA Arrays are not compatible with QCTool version 1.1.1.0.

If you are processing both GeneChip miRNA Arrays and GeneChip miRNA 2.0 Arrays you will be required to use a second computer because both versions of the miRNA QCTool cannot be installed on the same computer.

Conventions Used in This Guide

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

Steps

Instructions for procedures are written in a 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 that appears.

Font Styles

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

For example:

1. Click the **Find** button or select **Edit** → **Find** from the menu bar.
The Find dialog box appears.
2. Enter *AFFX-BioB-5_at* 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. The screen captures depicted in this guide may not exactly match the windows displayed on your screen.

Additional Comments



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



NOTE: The Note format presents important 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.

Resources

Documentation

This manual is available in Adobe Acrobat format (as *.pdf files) on the CD and is readable with the Adobe® Acrobat Reader® software, available at no charge from Adobe at <http://www.adobe.com>.

Technical Support

Affymetrix provides technical support to all licensed users via phone or E-mail. To contact Affymetrix® Technical Support:

Affymetrix, Inc.

3420 Central Expressway
Santa Clara, CA 95051 USA
E-mail: support@affymetrix.com
Tel: 1-888-362-2447 (1-888-DNA-CHIP)
Fax: 1-408-731-5441

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
E-mail: supporteurope@affymetrix.com
Tel: +44 (0) 1628 552550
Fax: +44 (0) 1628 552585

Affymetrix Japan, K. K.

ORIX Hamamatsucho Bldg, 7F
1-24-8 Hamamatsucho, Minato-ku
Tokyo 105-0013 Japan
Tel: +81-3-6430-4020
Fax: +81-3-6430-4021
salesjapan@affymetrix.com
supportjapan@affymetrix.com

miRNA QCTool Software Installation and First Steps

Installing the miRNA QCTool Software

Software Requirements

The miRNA QCTool software can be installed on the following operating systems:

- Microsoft Windows 2000 Professional with Service Pack 4.0 or higher
- Microsoft Windows XP with Service Pack 2.0 or higher
- Microsoft Windows Vista with Service Pack 1.0 or higher
- Microsoft.Net 2.0

Minimum Hardware Recommendations

The minimum hardware recommendations are:

- Memory (RAM): 1 GB (2 GB of RAM is recommended)
- Hard drive: 20 GB
- Processor: 2.0 GHz Intel Pentium or better

Software Installation

1. Install the software on a Microsoft Windows operating system by double clicking on the setup program.

The miRNA QCTool Setup window appears ([Figure 2.1](#)).

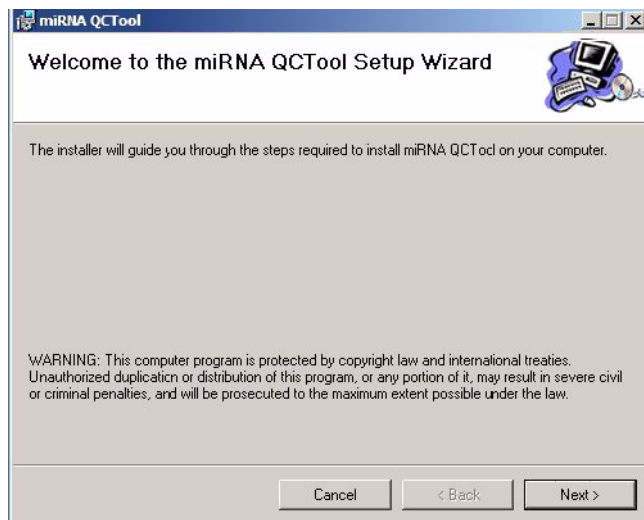


Figure 2.1 Setup Window

2. Click **Next**.

The miRNA QCTool License Agreement appears ([Figure 2.2](#)).



Figure 2.2 miRNA QCTool License Agreement

3. Read the License Agreement and select **I Agree**.
Selecting **I Do Not Agree** or **Cancel** will close the installer and miRNA QCTool will not be installed. Click **Next** to continue. The Select Installation Folder appears ([Figure 2.3](#)).

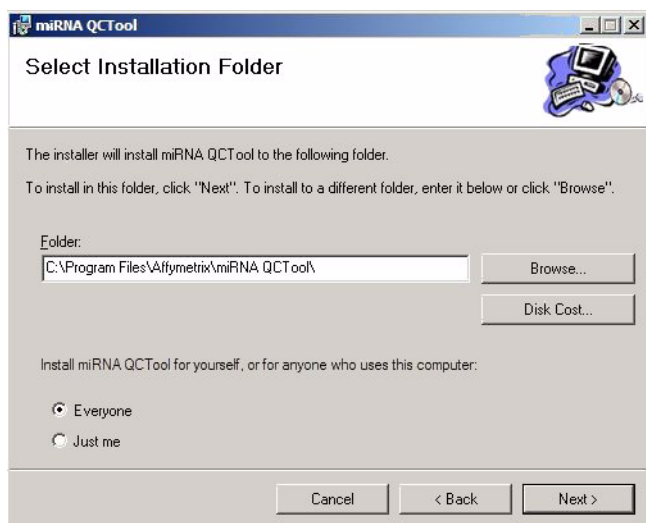


Figure 2.3 Select Installation Folder window

NOTE: The settings displayed in [Figure 2.3](#) show just one example of a location to install the software. You may install Affymetrix® miRNA QCTool software elsewhere, if desired.

4. In the Select Installation Folder window:
 - A. Click **Browse** to select the installation **Folder**.
 - B. Click **Everyone** or **Just me** to define who will be allowed to use the software.
 - C. Click **Next** to proceed.
The Installation Complete window appears ([Figure 2.4](#)).

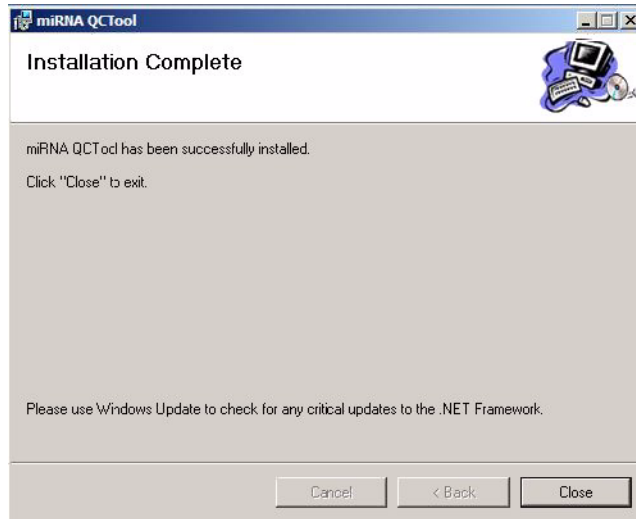


Figure 2.4 Installation Complete window

First Steps

Starting the miRNA QCTool Software

After installation of the miRNA QCTool program, launch the software by using one of two options:

- Click the desktop icon labeled “miRNA QCTool” or
- Launch the miRNA QCTool by clicking **Start** → **All Programs** → **Affymetrix** → **miRNA QCTool**.

The first operation is the loading of the necessary CEL files by using the CEL selection window (see [Selection of Input Data: CEL Files on page 11](#) for details). Once this step is accomplished, the software is able to process the raw data based on a pre-defined workflow of analysis steps. [Workflow Editing on page 23](#) explains how to configure, edit and save workflows.

The miRNA QCTool has been configured with the necessary information files (the default setup), so that the GC content for each probe and the list of background probes are automatically loaded upon starting. The section, [Software Settings](#) (below) explains how to reconfigure these settings.

Software Settings

The ability to change library file locations is available by accessing the **Tools** menu option and then clicking **Settings and Options...** A simple settings window allows the user to update the probe array information (see [Figure 2.5](#)), including the name and location of the CDF, annotation, and background files. The “default” files are the ones shipped with the program. The software settings include the **CDF File**, **Annotation File**, **Background BGP File**, **List of Probe Files** and **Quality Control QCC Files**.

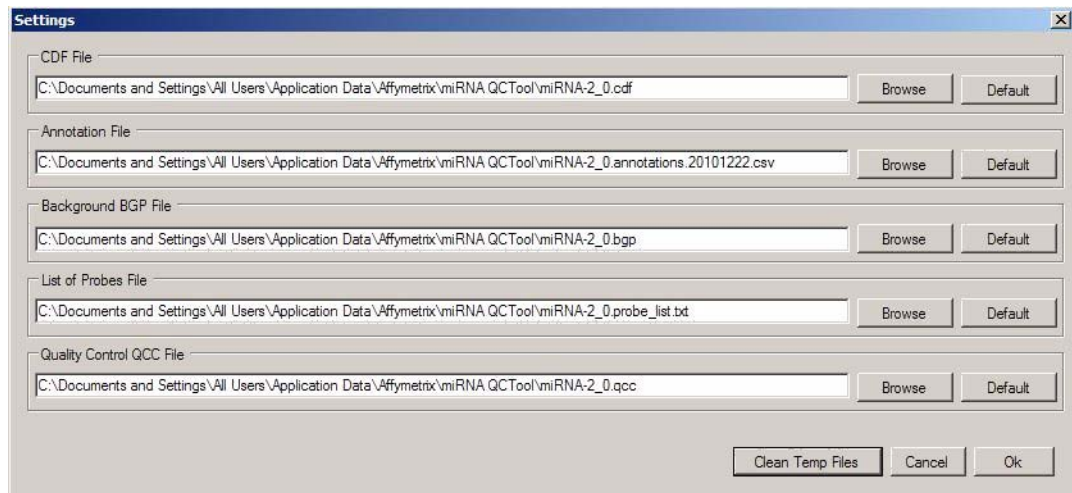


Figure 2.5 Settings window for updating configuration files

Window Arrangement

Windows arrangement is available through the **Window** menu option. This feature allows for the arrangement of open windows (tables and graphs), manually or automatically. Figure 2.6 and Figure 2.7 show two arrangement examples: (a) MvA and box plot graphs automatically in vertical tile mode, and (b) an MvA graph, a box plot graph and a data table manually in cascade window style.

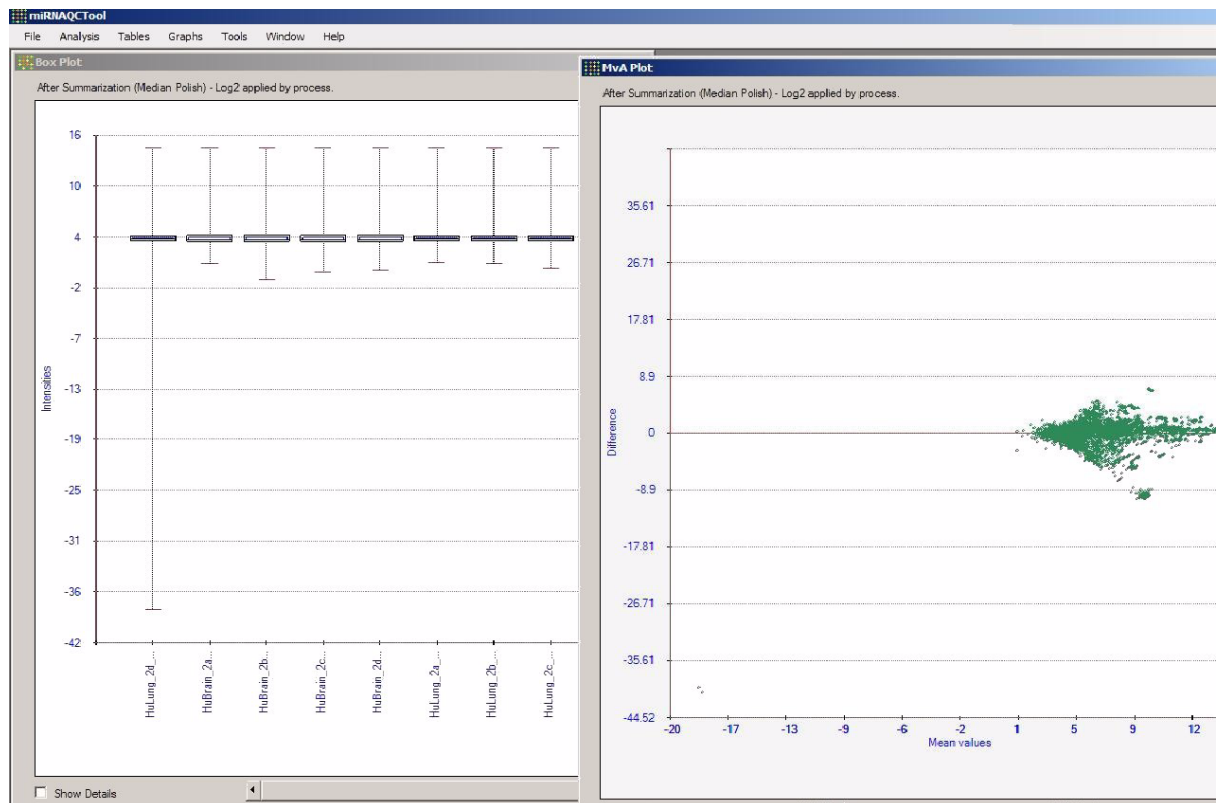


Figure 2.6 Example of multiple window arrangement: Vertical tile mode

Data after Summarization (Median Polish) - Log2 applied by process.

ProbeSet Name	HuLung_...	p-value (...)	Detectio...	HuBrain...	p-value (...)	Detectio...	HuBrain...	p-value (...)
cel-let-7_st	13.34865	8.721767E-16	True	13.0384312	8.721767E-16	True	12.8567171	8.721767E-16
cel-let-7-star_st	3.937564	0.6393688	False	3.606689	0.9783676	False	3.60258627	0.9987671
cel-hn-4_st	3.80536461	0.7640221	False	3.99884939	0.489643	False	3.92063522	0.4428738
cel-hn-4-star_st	4.151994	0.4638723	False	3.81604052	0.495932	False	3.76162362	0.7055392
cel-miR-1_st	4.04034138	0.7223458	False	4.12295055	0.4366831	False	3.971281	0.6307727
cel-miR-2_st	3.955419	0.656993	False	3.62169933	0.9314729	False	4.06695366	0.5779365
cel-miR-34_st	3.641423	0.7425691	False	3.55160761	0.5280683	False	3.78471231	0.3372566
cel-miR-34-star_st	4.720758	0.185002	False	5.16430426	0.04305392	True	5.196764	0.02023635
cel-miR-35_st	3.819388	0.9853876	False	3.82998085	0.8748435	False	3.60762239	0.9819251
cel-miR-36_st	4.18767166	0.1516112	False	4.19789028	0.1480861	False	3.973745	0.3414308
cel-miR-37-star_st	3.78736687	0.7755316	False	3.77236366	0.8254901	False	2.90533543	0.9989531
cel-miR-37_st	3.78736663	0.634245	False	4.0778923	0.4338754	False	3.97381258	0.5591431
cel-miR-38_st	4.214965	0.4738863	False	4.206868	0.04623421	True	4.20231771	0.1718871
cel-miR-39_st	4.10030651	0.9398	False	4.41063643	0.1544359	False	4.172182	0.2201207
cel-miR-40_st	4.298316	0.5550811	False	3.601266	0.8854913	False	3.902637	0.7962239
cel-miR-41_st	3.69530344	0.8738484	False	3.52844167	0.881606	False	3.77539182	0.5585493
cel-miR-42-star_st	3.40979528	0.9513109	False	3.601266	0.8571414	False	3.81006	0.4592267
cel-miR-42_st	3.709423	0.8501645	False	3.98974943	0.7284735	False	3.820654	0.9069499
cel-miR-43_st	4.03508854	0.5695897	False	4.07139969	0.2019512	False	4.173333	0.3269734

Showing data after Summarization (Median Polish) - Log2 applied by process.

Figure 2.7 Example of multiple window arrangement: Cascade arrangement

Selection of Input Data: CEL Files

Each time the software starts, a CEL-files selection window opens up. This window contains a list of previously selected files. On the first run, the CEL selection window (Figure 2.8) may appear empty.

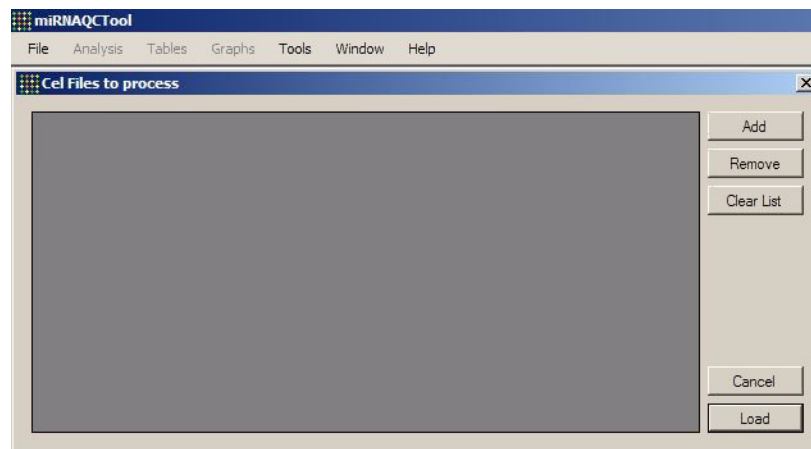


Figure 2.8 CEL Selection window

The addition of CEL files is accomplished by clicking the **Add** button, and selecting the CEL files through a standard Windows input dialog (which allows for multiple selections). These files can be removed from the list, using the **Remove** or **Clear List** button. Once all the CEL files needed for the analysis are selected, click the **Load** button. When finished, the selection window closes, and you are notified that the cel files were loaded successfully. The **Analysis** menu is enabled.

CEL files can be loaded later by accessing the **File** menu. More details are provided in [Chapter 3, File Menu](#) on page 13.

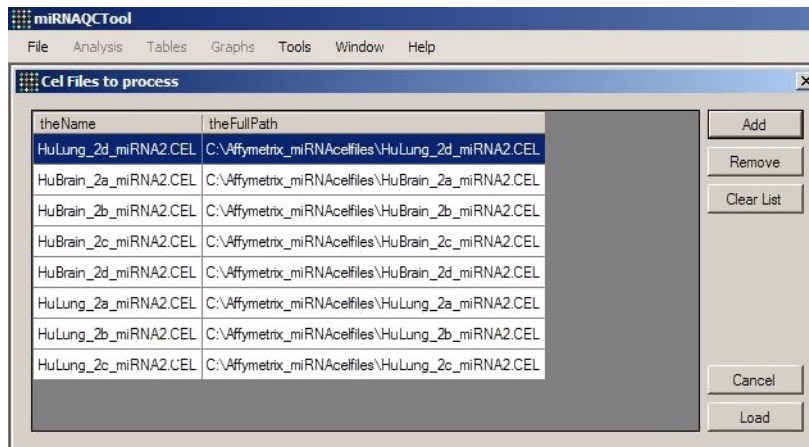


Figure 2.9 The interface for CEL files input. It displays the CEL file names and the full path. It allows for the adding of files one at a time, or by selecting all the files together. Mouse hover shows a tooltip with the full file name.

Menus

File Menu

The basic input/output management is handled by the File menu (see [Figure 3.1](#)). This section describes in detail each option of this menu.

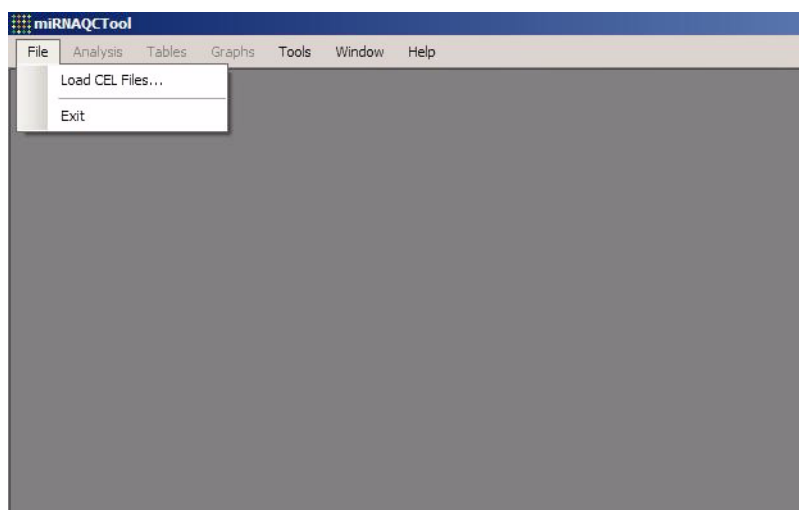


Figure 3.1 File Menu

Loading CEL Files

To load CEL files, choose **Load CEL Files...** from the **File** menu. The **CEL files to process** window appears. The use of this window is described in the section [Selection of Input Data: CEL Files on page 11](#).

Tables Menu

Tables are available for each point of the analysis, as defined in the workflow¹. Click the menu option **Tables** (see [Figure 3.2](#)) to see the available options. All the tables can be exported and saved.

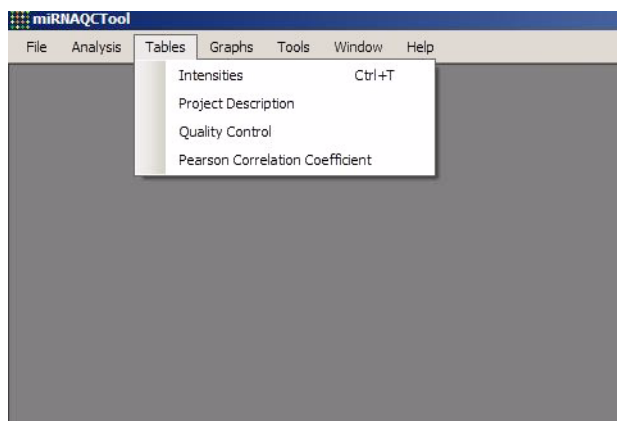
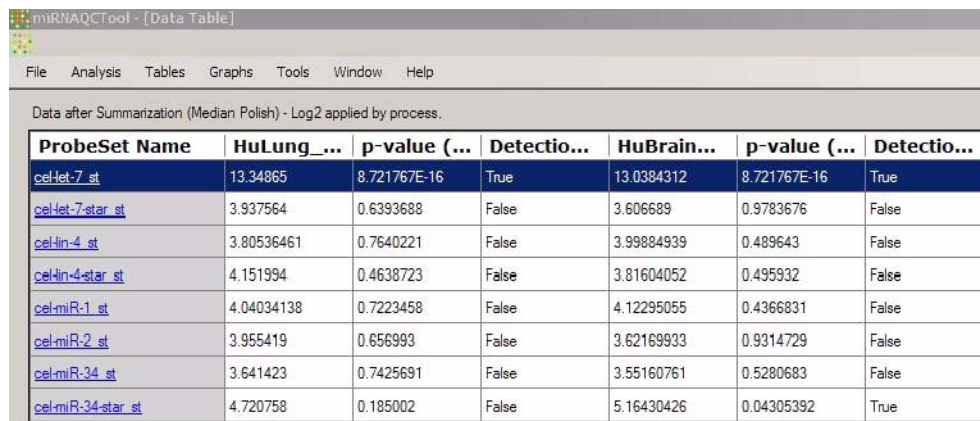


Figure 3.2 Tables Menu Options

¹ Prior to accessing this feature, an analysis must be run or this menu option will appear shadowed.

Data Tables / Intensities

Data tables are provided for each step in the analysis workflow. [Figure 3.3](#) and [Figure 3.4](#) show examples of this window. The tables show the intensities in a matrix format, where columns correspond to CEL files and rows correspond to probes or probe sets (depending on the stage being visualized).



miRNAQCTool - [Data Table]

File Analysis Tables Graphs Tools Window Help

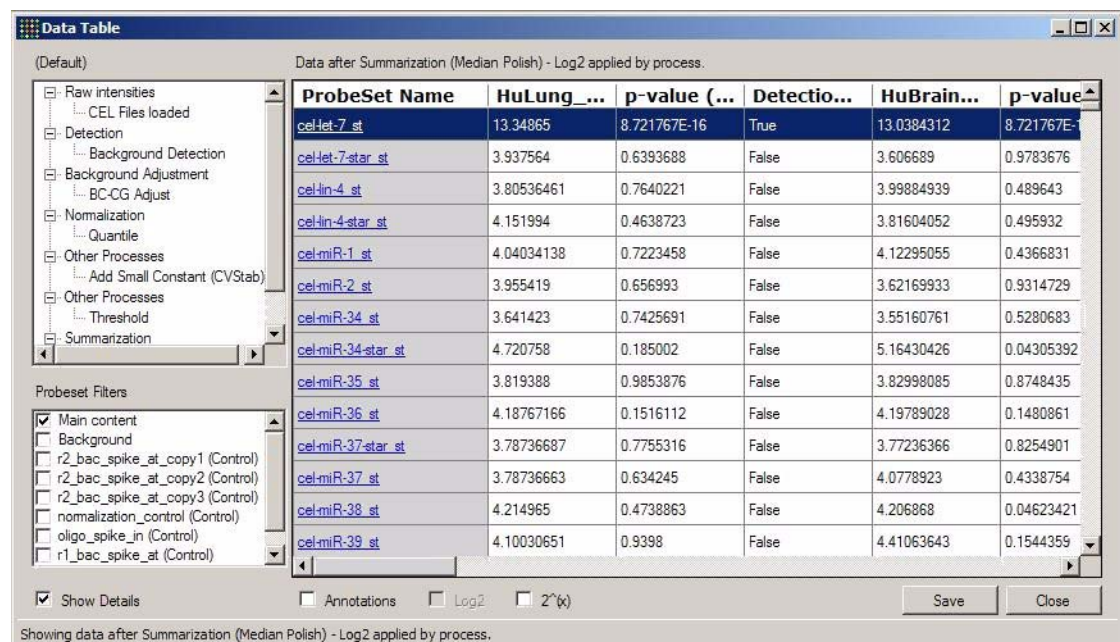
Data after Summarization (Median Polish) - Log2 applied by process.

ProbeSet Name	HuLung_...	p-value (...)	Detectio...	HuBrain...	p-value (...)	Detectio...
celHet-7_st	13.34865	8.721767E-16	True	13.0384312	8.721767E-16	True
celHet-7-star_st	3.937564	0.6393688	False	3.606689	0.9783676	False
celIn-4_st	3.80536461	0.7640221	False	3.99884939	0.489643	False
celIn-4-star_st	4.151994	0.4638723	False	3.81604052	0.495932	False
cel-miR-1_st	4.04034138	0.7223458	False	4.12295055	0.4366831	False
cel-miR-2_st	3.955419	0.656993	False	3.62169933	0.9314729	False
cel-miR-34_st	3.641423	0.7425691	False	3.55160761	0.5280683	False
cel-miR-34-star_st	4.720758	0.185002	False	5.16430426	0.04305392	True

Figure 3.3 Example of table displaying the summarized data

[Figure 3.3](#) and [Figure 3.4](#) show an example of the analysis result displayed as a data table. In [Figure 3.4](#), the **Show Details** checkbox was selected to access the workflow and filters.

- The **Probeset Filters** checkboxes allow you to add or remove content from the table, such as main content, control, background and other probes present in the chip.
- The data can be displayed as is or after **Log₂** and **2^(x)** transforms.
- Selection of the **Annotations** option adds annotation to the table.
- Tables can be exported as text files by clicking the **Save** button.



Data Table

(Default) Data after Summarization (Median Polish) - Log2 applied by process.

Raw intensities
... CEL Files loaded

Detection
... Background Detection

Background Adjustment
... BC-CG Adjust

Normalization
... Quantile

Other Processes
... Add Small Constant (CVStab)

Other Processes
... Threshold

Summarization

Probeset Filters

Main content

Background

r2_bac_spike_at_copy1 (Control)

r2_bac_spike_at_copy2 (Control)

r2_bac_spike_at_copy3 (Control)

normalization_control (Control)

oligo_spike_in (Control)

r1_bac_spike_at (Control)

ProbeSet Name	HuLung_...	p-value (...)	Detectio...	HuBrain...	p-value...
celHet-7_st	13.34865	8.721767E-16	True	13.0384312	8.721767E-16
celHet-7-star_st	3.937564	0.6393688	False	3.606689	0.9783676
celIn-4_st	3.80536461	0.7640221	False	3.99884939	0.489643
celIn-4-star_st	4.151994	0.4638723	False	3.81604052	0.495932
cel-miR-1_st	4.04034138	0.7223458	False	4.12295055	0.4366831
cel-miR-2_st	3.955419	0.656993	False	3.62169933	0.9314729
cel-miR-34_st	3.641423	0.7425691	False	3.55160761	0.5280683
cel-miR-34-star_st	4.720758	0.185002	False	5.16430426	0.04305392
cel-miR-35_st	3.819388	0.9853876	False	3.82998085	0.8748435
cel-miR-36_st	4.18767166	0.1516112	False	4.19789028	0.1480861
cel-miR-37-star_st	3.78736687	0.7755316	False	3.77236366	0.8254901
cel-miR-37_st	3.78736663	0.634245	False	4.0778923	0.4338754
cel-miR-38_st	4.214965	0.4738863	False	4.206868	0.04623421
cel-miR-39_st	4.10030651	0.9398	False	4.41063643	0.1544359

Show Details Annotations Log₂ 2^(x)

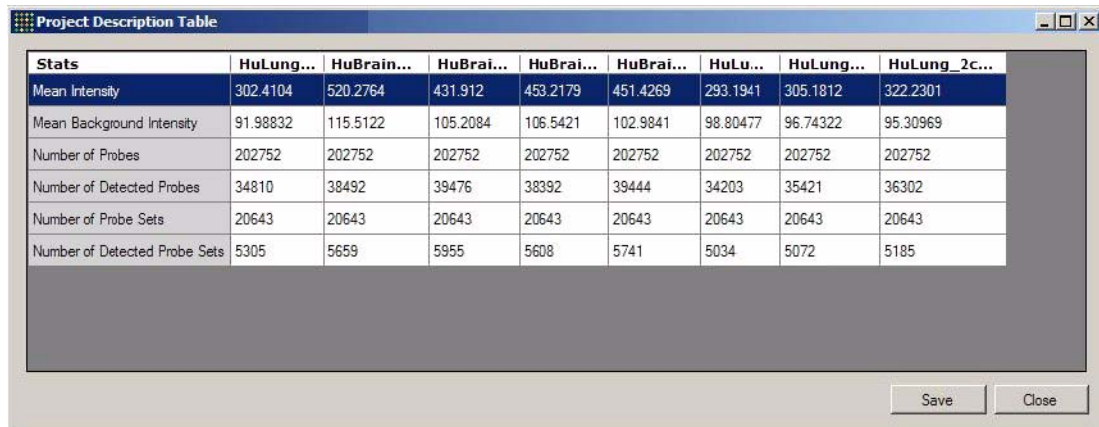
Save Close

Showing data after Summarization (Median Polish) - Log2 applied by process.

Figure 3.4 Example of table displaying the workflow and filters

Project Description Table

This table displays details of the current analysis. Figure [Figure 3.5](#) shows an example of the table using eight CEL files.

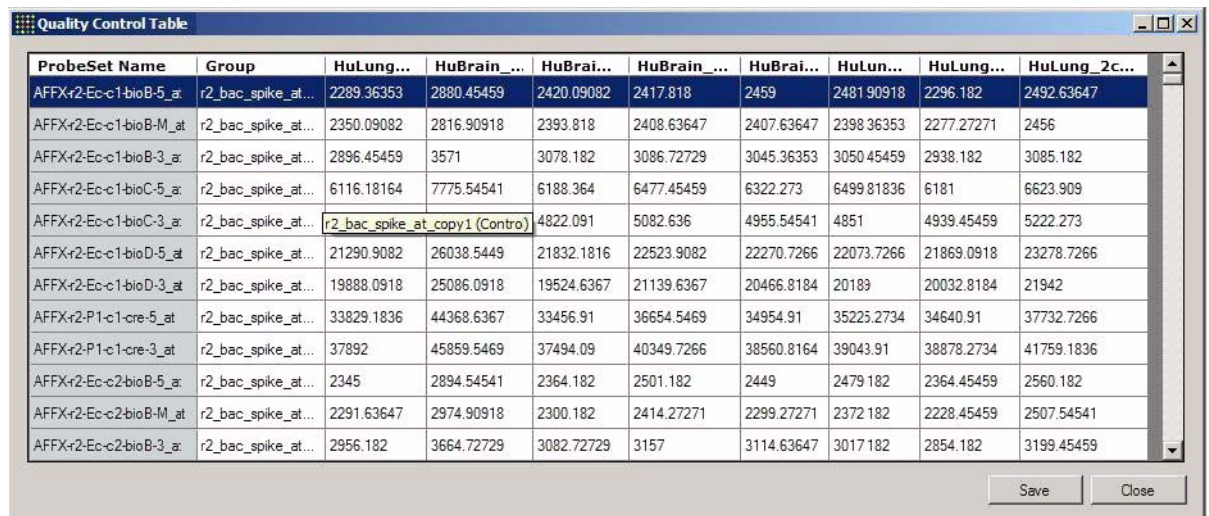


Stats	HuLung...	HuBrain...	HuBrai...	HuBrai...	HuBrai...	HuLu...	HuLung...	HuLung_2c...
Mean Intensity	302.4104	520.2764	431.912	453.2179	451.4269	293.1941	305.1812	322.2301
Mean Background Intensity	91.98832	115.5122	105.2084	106.5421	102.9841	98.80477	96.74322	95.30969
Number of Probes	202752	202752	202752	202752	202752	202752	202752	202752
Number of Detected Probes	34810	38492	39476	38392	39444	34203	35421	36302
Number of Probe Sets	20643	20643	20643	20643	20643	20643	20643	20643
Number of Detected Probe Sets	5305	5659	5955	5608	5741	5034	5072	5185

Figure 3.5 Example of table displaying the workflow and filters

Quality Control Table

This table displays chip-specific control probes. For each CEL file, the values displayed are the average (over the probes) raw intensities for each probe set. [Figure 3.6](#) shows an example of the table using eight CEL files.

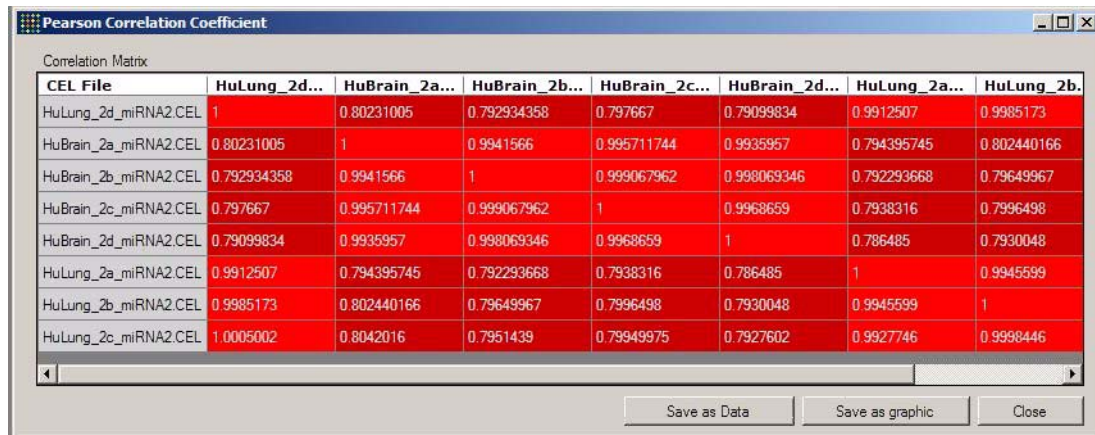


ProbeSet Name	Group	HuLung...	HuBrain...	HuBrai...	HuBrain...	HuBrai...	HuLun...	HuLung...	HuLung_2c...
AFFX-r2-Ec-c1-bioB-5_at	r2_bac_spike_at...	2289.36353	2880.45459	2420.09082	2417.818	2459	2481.90918	2296.182	2492.63647
AFFX-r2-Ec-c1-bioB-M_at	r2_bac_spike_at...	2350.09082	2816.90918	2393.818	2408.63647	2407.63647	2398.36353	2277.27271	2456
AFFX-r2-Ec-c1-bioB-3_at	r2_bac_spike_at...	2896.45459	3571	3078.182	3086.72729	3045.36353	3050.45459	2938.182	3085.182
AFFX-r2-Ec-c1-bioC-5_at	r2_bac_spike_at...	6116.18164	7775.54541	6188.364	6477.45459	6322.273	6499.81836	6181	6623.909
AFFX-r2-Ec-c1-bioC-3_at	r2_bac_spike_at...	r2_bac_spike_at_copy1 (Contro)	4822.091	5082.636	4955.54541	4851	4939.45459	5222.273	
AFFX-r2-Ec-c1-bioD-5_at	r2_bac_spike_at...	21290.9082	26038.5449	21832.1816	22523.9082	22270.7266	22073.7266	21869.0918	23278.7266
AFFX-r2-Ec-c1-bioD-3_at	r2_bac_spike_at...	19888.0918	25086.0918	19524.6367	21139.6367	20466.8184	20189	20032.8184	21942
AFFX-r2-P1-c1-cre-5_at	r2_bac_spike_at...	33829.1836	44368.6367	33456.91	36654.5469	34954.91	35225.2734	34640.91	37732.7266
AFFX-r2-P1-c1-cre-3_at	r2_bac_spike_at...	37892	45859.5469	37494.09	40349.7266	38560.8164	39043.91	38878.2734	41759.1836
AFFX-r2-Ec-c2-bioB-5_at	r2_bac_spike_at...	2345	2894.54541	2364.182	2501.182	2449	2479.182	2364.45459	2560.182
AFFX-r2-Ec-c2-bioB-M_at	r2_bac_spike_at...	2291.63647	2974.90918	2300.182	2414.27271	2299.27271	2372.182	2228.45459	2507.54541
AFFX-r2-Ec-c2-bioB-3_at	r2_bac_spike_at...	2956.182	3664.72729	3082.72729	3157	3114.63647	3017.182	2854.182	3199.45459

Figure 3.6 Quality Control Table

Pearson Correlation Coefficient Table

This table displays the Pearson Correlation Coefficient between CEL files (Figure 3.7).



The screenshot shows a window titled "Pearson Correlation Coefficient" with a "Correlation Matrix" table. The table has 8 columns and 8 rows, with the diagonal elements all equal to 1. The columns are labeled with CEL file names: HuLung_2d..., HuBrain_2a..., HuBrain_2b..., HuBrain_2c..., HuBrain_2d..., HuLung_2a..., HuLung_2b..., and HuLung_2c... The rows are labeled with the same file names. The values represent the Pearson correlation coefficients between the files.

CEL File	HuLung_2d...	HuBrain_2a...	HuBrain_2b...	HuBrain_2c...	HuBrain_2d...	HuLung_2a...	HuLung_2b...
HuLung_2d_miRNA2.CEL	1	0.80231005	0.792934358	0.797667	0.79099834	0.9912507	0.9985173
HuBrain_2a_miRNA2.CEL	0.80231005	1	0.9941566	0.995711744	0.9935957	0.794395745	0.802440166
HuBrain_2b_miRNA2.CEL	0.792934358	0.9941566	1	0.999067962	0.998069346	0.792293668	0.79649967
HuBrain_2c_miRNA2.CEL	0.797667	0.995711744	0.999067962	1	0.9968659	0.7938316	0.7996498
HuBrain_2d_miRNA2.CEL	0.79099834	0.9935957	0.998069346	0.9968659	1	0.786485	0.7930048
HuLung_2a_miRNA2.CEL	0.9912507	0.794395745	0.792293668	0.7938316	0.786485	1	0.9945599
HuLung_2b_miRNA2.CEL	0.9985173	0.802440166	0.79649967	0.7996498	0.7930048	0.9945599	1
HuLung_2c_miRNA2.CEL	1.0005002	0.8042016	0.7951439	0.79949975	0.7927602	0.9927746	0.9998446

Figure 3.7 Pearson Correlation Coefficient

Graphs Menu

The miRNA QCTool software provides several ways of displaying results for analysis.

- *Box Plots*
- *Histogram Plots*
- *MvA Plots*
- *Quality Control Plots*

Select the **Graphs** menu option to see the available options (see Figure 3.8).

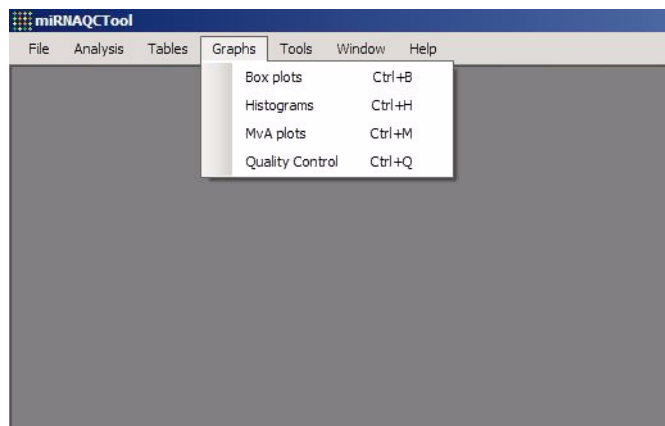


Figure 3.8 Graphs Menu Options

For all plots, similar to the **Tables** window, the user can choose to display graphical results for any analysis stage by selecting the desired one from the workflow view on the left side of the window when they click **Show Details**.

It is possible to save the graphs to various image formats, bitmap (bmp), Jpeg (jpg), and others, by clicking on the **Save** button.

Box Plots

The box-whisker plots are used to evaluate the overall consistency of the different samples within the experiment. The plot summarizes the distribution of miRNA expression values for each sample so one can readily identify potential outliers with distributions that may be different from the other samples in the project. One can also evaluate the effect of normalization and how that influences the consistency of distributions across the entire experiment. Selecting the **Box Plot** menu option displays a box plot of all the CEL files, and allows zooming in by selecting the number of whiskers. [Figure 3.9](#) shows an example of the Box Plot window.

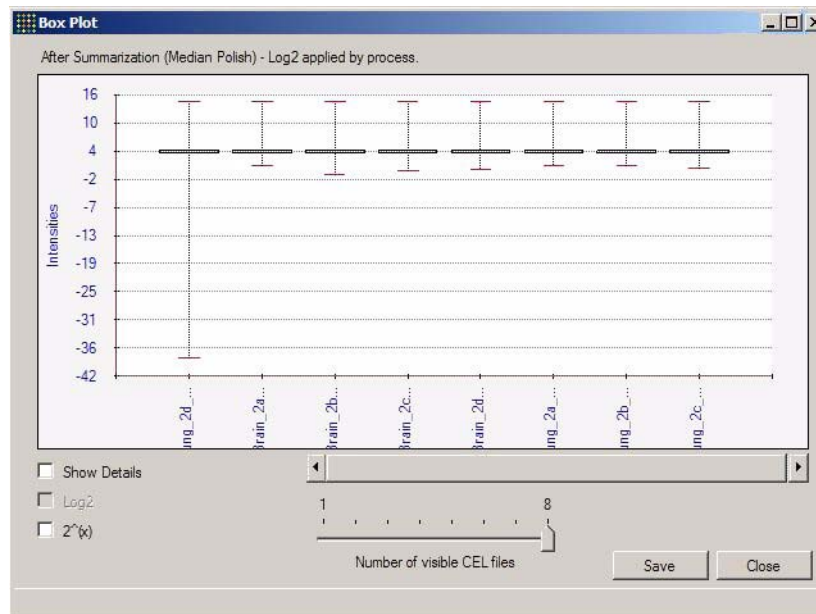


Figure 3.9 Sample of box plot image for some sample CEL files

Main properties of the box plots are listed below:

- The line in the box indicates the median value of the data.
- The upper edge of the box marks the 75th percentile of the expression values.
- The lower edge of the box indicates the 25th percentile.
- The distance between the two quartiles is known as the inter-quartile range.
- The ends of the lines (whiskers) indicate the minimum and maximum data values.

The graphic, by default, includes all the chip probes types: background, miRNA, control and snoRNA. The user can select which ones to include by checking or unchecking the checkboxes (visible when checking **Show Details**).

Moving the mouse over a box plot displays the CEL file name, minimum, median and maximum intensities. The data can be displayed as is or after **Log2** and **2(x)** transforms.

Histogram Plots

The user can access this feature through the **Histogram** option on the **Graphs** menu. The histogram image (see [Figure 3.10](#)) displays one or more CEL files in the same graph, by selecting them from the top right list. It shows the original data unless the Log2 checkbox is selected. The histogram can be normalized to show a density function. The smoothing option displays a smoother histogram.

Another option allows the user to display an aggregated histogram of many CEL files, selected in the second list.

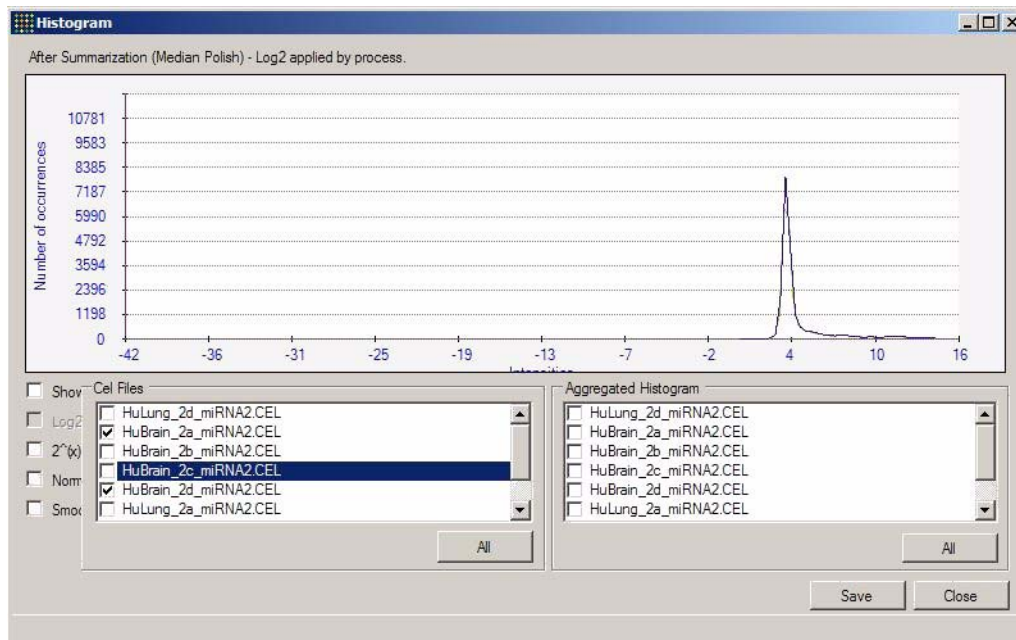


Figure 3.10 Sample of histogram image of some sample CEL files

MvA Plots

The MvA plot compares two experimental groups of samples. It shows the changes in expression between two groups as a function of the average expression level of both groups. Each spot represents a unique product (miRNA, snoRNA) as probe or probe set, depending on the analysis stage. The vertical axis is the log₂ fold-change, computed as the difference between the means (M) of the two samples¹. The horizontal axis shows the mean normalized signal (A) computed as the average of both samples (CEL files). The user can access this feature through the **MvA Plots** option on the **Graphs** menu. [Figure 3.11](#) shows an MvA plot window.

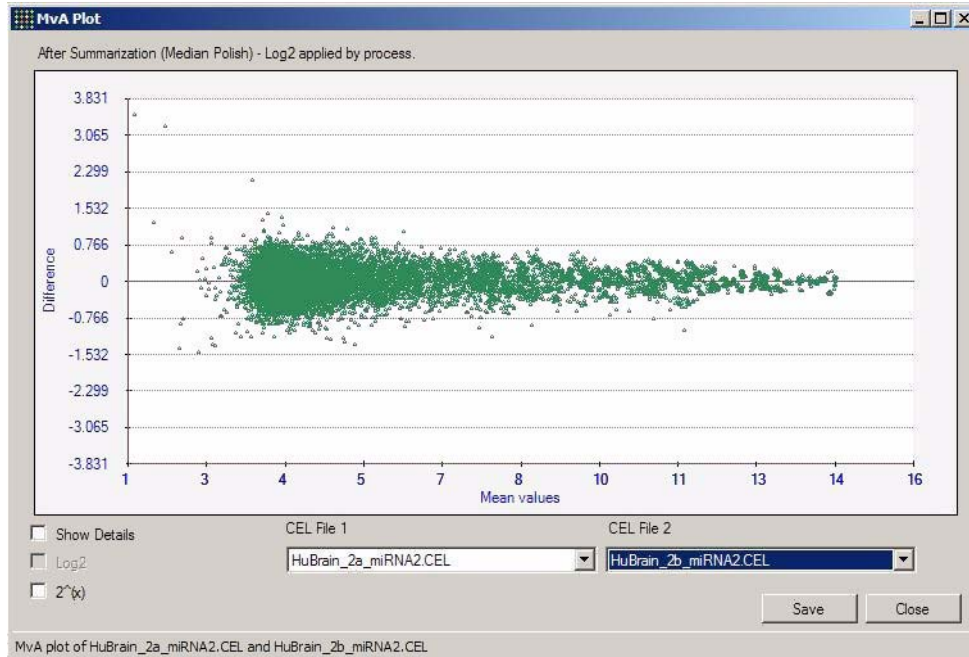


Figure 3.11 Sample of MvA plot for two CEL files

The MvA window displays the following information:

- The analysis stage used to generate the MvA plot.
- The two CEL files compared in the MvA plot.

These plots enable one to observe fundamental distortions in the data that may exist between the two experimental groups. For example, if two experimental samples are well-matched, the majority of spots will cluster along the horizontal line where $y = 0$. Furthermore, the extent to which the spots spread out horizontally is an indicator of the dynamic range of those samples. If one sample replicate is plotted against another single sample, it shows the variability as a function of the mean value for the two arrays.

¹ For instance, if the vertical axis is 2, there is approximately a four-fold difference between the expression values of the two samples.

Quality Control Plots

The Quality Control plot shows the average intensity (over the probes) of the quality probes sets across the CEL files.

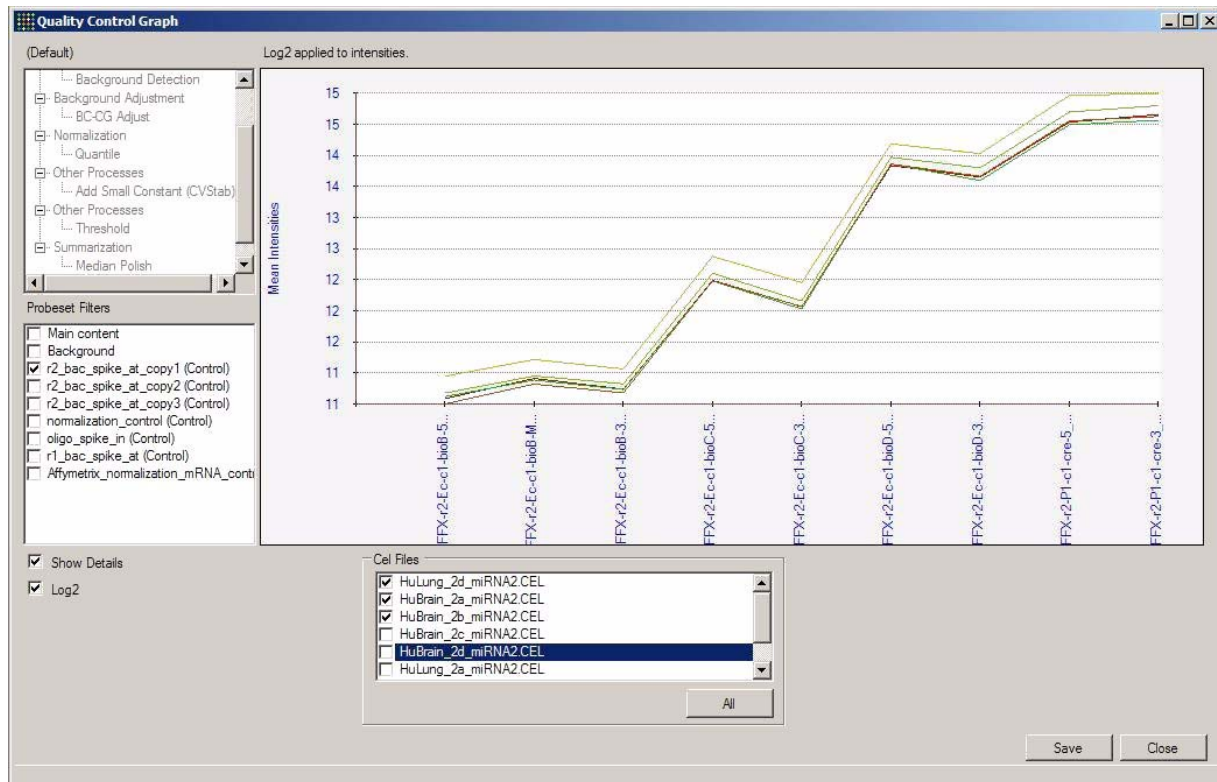


Figure 3.12 Sample of Quality Control Plot

Analysis

Processing Steps

The processing algorithms are briefly described here. A more complete description of the algorithms can be found in [Appendix A, Algorithms on page 25](#).

Probe specific signal detection calls: Each probe on an array is assayed for detection.

For miRNA probes, detection is based on a Wilcoxon Rank-Sum test of the miRNA probe set signals compared to the distribution of signals from GC content matched anti-genomic probes. If the resulting p-value for the probes is $p \leq 0.06$, it is considered “detected above background”. For non-miRNA probes, detection is at probe level, where for each probe, the p-value is the probability of generating a value equal to or greater than the intensity, from the distribution of the background probes. Probe set p-value is computed using a Fisher test, as a percentage of the probe's p-values (default is 50% or the median). Probes with p-values > 0.06 have insufficient signal to discriminate from the background and are thus considered “not detected”.

Background estimation and correction: The same set of anti-genomic probes used to determine detection calls are used to estimate GC content matched background signals. Each miRNA probe signal has a GC content matched background estimate subtracted from its value. This GC-specific background contribution is estimated by the median signal from the distribution of GC-matched anti-genomic probes. The resulting value may be negative at this stage, if the probe's signal is less than the median value of the GC-matched anti-genomic probes. The same process can be applied using two other ways to match main content probes to background probes: a) by GC percent and b) by both GC content and probe length.

Constant Variance Stabilization on probes: Probe level constant variance stabilization is based on adding a small constant to all the intensities. The default value of this constant is 16.

Normalization: Probe level normalization is provided via quantile, mean array scaling, and mean array scaling (based on normalization probes) normalization algorithms.

Summarization: This process provides the quantification of probe intensities into probe set intensities (via mean, median or RMA methods).

BC-CG Adjust: Background adjustment using ‘anti-genomic’¹ background probes, matched to miRNA array species probe by GC count.

BC-CG % Adjust: Background adjustment using ‘anti-genomic’ background probes, matched to miRNA array species probe by GC percentage.

BC-CG Adjust and Length adjust: Background adjustment using ‘anti-genomic’ background probes, matched to miRNA array species probe by both length and GC count.

Quantile normalization: To compare the probe intensities across chips, cross chip variation needs to be addressed, and the values normalized. A very commonly used and accepted normalization method that assures that the distribution of log (probe intensities) is comparable among experiments, see BM Bolstad RA Irizarry M Anstrand & TP Speed (2003) “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”. *Bioinformatics*, 19 (2):185-193.

Mean array Scaling: A simple normalization method that assures that the mean of the log (intensities) of probes are the same across experiments.

Mean array scaling (on Normalization probes): Separately apply mean scaling to normalization probes.

Add small Constant: To avoid probe intensities to drop below zero after Background-GC correction a small constant is added to all intensities.

Threshold: Negative values are replaced by zero.

Mean: Intensities for probe sets are derived by taking the mean of the normalized intensities within the probe set.

¹ Sequences not found in human, mouse or rat genomes.

Median polish: Median polish is a summarization method that essentially ignores outlier probe-level values. *Biostatistics*. 2003 Apr;4(2):249-64. 2003.

Median: Intensities for probe sets are derived by taking the median of the normalized intensities within the probe set.

Running an Analysis

To run or edit an analysis workflow, choose the **Analysis** menu option on the menu bar. The user is presented with a single window displaying the selected workflow (see [Figure 4.1](#)), allowing for flexibility regarding the addition of new options. In the event that the user wants to use the selected workflow with default parameters, one just needs to click on the **Run** button. To select a pre-defined or a custom workflow, the user can click the **Workflows** button, and choose it from a list of pre-defined workflows.

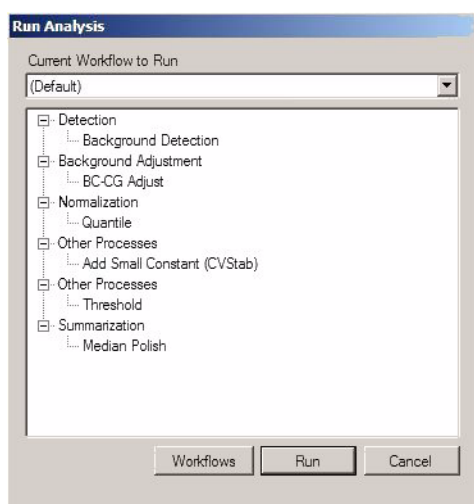


Figure 4.1 Interface for Running an Analysis

The default parameters for each option can be modified by clicking **File** → **New** or **File** → **New from Current**, or by accessing the Tools menu, and then selecting **Workflow Editor**. (See [Figure 4.2](#)). The following section provides more details about editing workflows.

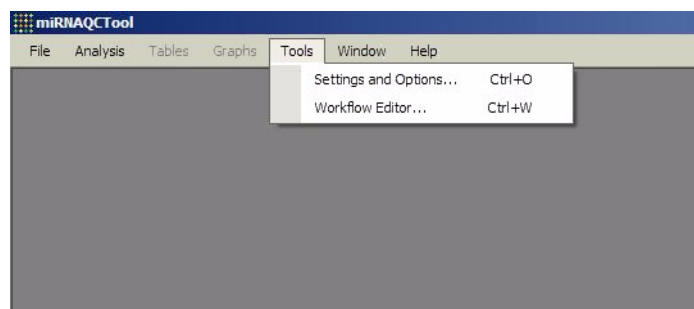






Figure 4.2 Tools menu option to edit workflows

Workflow Editing

The workflow editor (see [Figure 4.3](#)) allows the user to add and modify analysis workflows.

The miRNA QCTool comes with pre-defined workflows, which provide the basic analysis steps, with default parameters. If the user is comfortable with these, there is no need to access the workflow editing part of the program, and one can simply select a pre-defined workflow with the **Run** button when running an analysis. If the user wants to define specific analysis paths, with different options and/or parameters, one can create them and save them using the workflow editor.

[Figure 4.3](#) shows the main window for the creation and editing of workflows. The workflow tree on the left side shows all the **Available Processes**. The workflow tree on the right side shows a proposed workflow, built from processes listed on the left side.

- To create a custom workflow, click **File** → **New** or **File** → **New from Current**. Give a name to the custom workflow and click **OK**.
- To add a step, simply click to select one of the available options in the left window, then click the middle arrow button .
- To eliminate a step, click to select it in the right window, then click the  button to remove it.
- To change the order of the analysis, click to select the step in the right window, then click the up  or down  button.

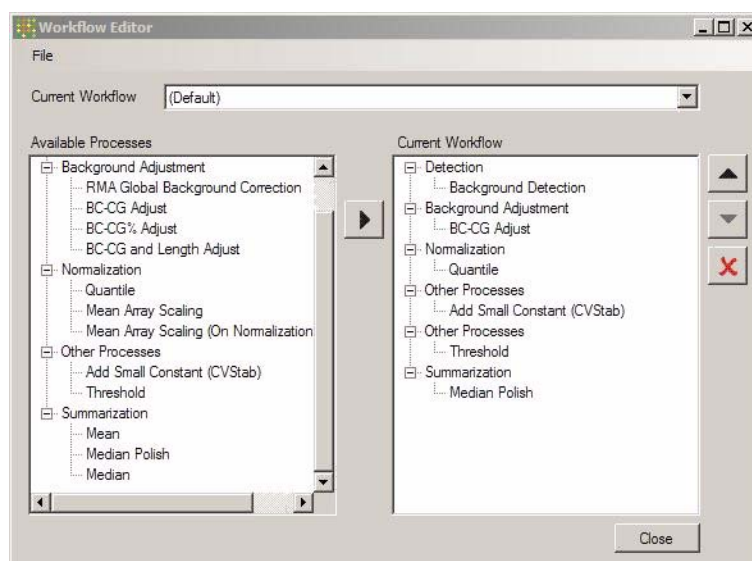


Figure 4.3 Interface for workflow edition

After you have finished creating your new workflow, click **File** → **Save** (See [Figure 4.4](#)). A dialog box appears indicating that “The workflow was updated.”

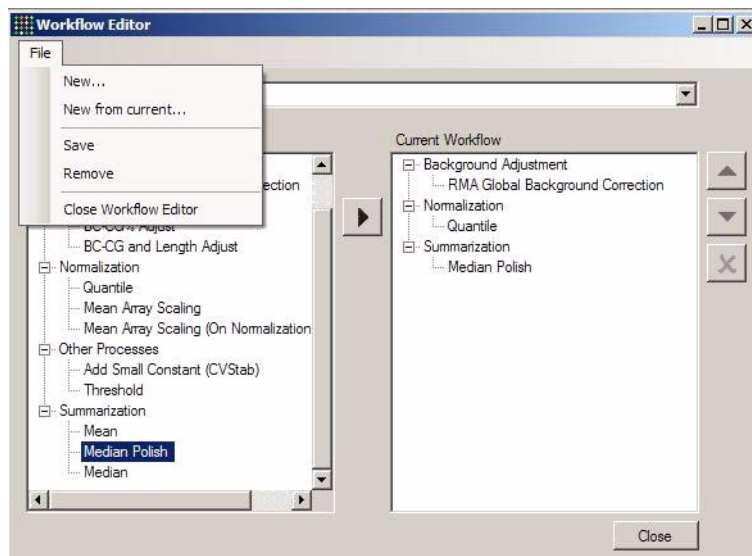


Figure 4.4 Saving the new workflow

All basic transformations, including detection for non miRNA probes, background adjustment, and normalization, are implemented at probe level. Detection for miRNA probes is done at the probe set level. Summarization is the operation that converts probe level data into probe set level data.

The following is a list of data processing algorithms that can be included in the workflow.

1. Detection
 - A. DABG - Detected Above Background (using Wilcoxon Test for miRNAs)
2. Background Adjustment
 - A. BC-GC Adjust - Adjustment by GC content
 - B. BC-GC% Adjust - Adjustment by GC percent
 - C. BC-GC and Length Adjust - Adjustment by GC content and Probe Length
 - D. RMA background adjustment
3. Data Transformation
 - A. CVStab - Constant Variance Stabilization (Intensity + 16)
 - B. Threshold - Thresholding for negative values
4. Normalization
 - A. Quantile Normalization
 - B. Mean Array Scaling Normalization
 - C. Mean Array Scaling (on normalization probes) Normalization
5. Summarization
 - A. Median Polish - Median Polish Summarization
 - B. Mean - Mean Value Summarization
 - C. Median - Median Value Summarization

Algorithms

Detection

The detection algorithm is similar to the one implemented in the Affymetrix Power Tools (APT)¹ Software, for all non-replicated probe sets, and uses the Wilcoxon Rank-Sum test for replicated probe sets compared to the distribution of signals from GC content matched antigenomic probes. Replicated probe sets are probe sets where all probes are identical replicates, and contain the same sequence. For this chip, replicated probe sets are miRNA probe sets.

Detection Algorithm Description

The detection process is described below. The algorithm is applied independently on each CEL file and probe set:

1. Process each CEL file one at a time.
2. The background probes intensities are grouped by GC content, into “Bins”, from 1 to 25 (Bins 1, 2, 3 have no probes).
3. For each probe set, software determines if it corresponds to a miRNA or another type.
4. If the probe set corresponds to a miRNA, the Wilcoxon test is applied. If not, the standard APT test is applied.
5. Both tests are applied to one probe set (and its probes) at a time.

Wilcoxon Test

1. All the probes of the probe set are assumed to have the same sequence and same GC content.
2. Program determines the GC content for the probe set based on the GC content of the first probe.
3. Program tests the probes raw intensities against all the anti-genomic probes associated with the same GC content (usually 96 probe sets, times 4 probes, equals 384 values), using the one-sided Wilcoxon Rank-Sum non-paired test.
4. Usually, the two vectors to be compared are:
 - A. vector with 4 intensity values for the probe set being tested, and
 - B. vector with approximately 384 intensity values for the anti-genomic probes with same GC content.
5. The p-value obtained from the test is assigned to both the probe set and all of its associated probes.

Affymetrix Test

1. For each probe within the probe set, the software determines its GC content and then looks at the distribution of the background probes for the same CEL file and GC content.
2. The p-value is defined by the probability of generating a value equal to or greater than the intensity of the probe, based on the distribution of the background probes.
3. **For both probes and probe sets, the Flag is defined as “1” when the p-value is less than or equal to 0.06, and “0” otherwise. Flag=1 indicate “detected” probes and/or probe sets.**
4. This process is repeated for each CEL file.

¹ Affymetrix Power Tools are a set of cross-platform command line programs that implement algorithms for analyzing and working with Affymetrix GeneChip arrays. APT is an open-source project licensed under the GNU General Public License (GPL).

Background Adjust

The Background Adjust Algorithm is a new implementation of GC content background subtraction.

Algorithm Description

For each CEL file:

1. The background probes intensities are grouped by GC content, into “Bins”, from 1 to 25 (Bins 1, 2 and 3 have no data).
2. For each Bin, all the intensities of the probes with the associated GC content are used to compute a median intensity.
3. For Bins with no data, use the median intensity of Bin 12.
4. The median for each Bin is the “correction” value for its associated GC bin.
5. For each non-background probe in the dataset, the program determines the GC content, and its correction value, using the form probe signal- BG correction.

Wilcoxon Rank-Sum Test for Detection

Replicated miRNA probes are assayed for detection based on a Wilcoxon Rank-Sum test of the miRNA probe signal compared to the distribution of signals from GC content matched anti-genomic probes. The Wilcoxon Rank-Sum test compares two distributions to assess whether one has systematically larger values than the other, based on the rank sum statistic W . For two samples of size n_1 and n_2 , it computes the rank sum statistic w as the sum of the index (or rank) of the elements of the first sample after sorting all the $n_1 + n_2$ values as a unique list. The test is based on the null hypothesis that the two samples are produced by the same distribution. From the distribution of rank sums, when the null hypothesis is true, a p -value is computed. If the p -value is small, the null hypothesis is rejected, and the probe is considered present. The distribution of the rank sum statistic w becomes approximately normal when the two samples are large. In this case the new variable:

$$Z = (W - \mu_w) / \sigma_w \text{ (A.1)}$$

is a z -statistic, where:

$$\mu_w = n_1 * (n_1 + n_2 - 1) / 2 \text{ (A.2)}$$

and

$$\sigma_w = \sqrt{(n_1 * n_2 * (n_1 + n_2 + 1) / 12)} \text{ (A.3)}$$

In this case, the test is approximated by $p = P(Z \geq z)$, where z is the standardized version of the computed rank sum W . **We define a probe as “detected” if its p -value $P(Z \geq z)$ is lower or equal to 0.06.**

Algorithm Description

1. Let n_1 (usually 4) be the number of non-background probes associated with a miRNA probe set, and n_2 the number of background probes with the same GC content.
2. Rank the n_1 probes in the $n_1 + n_2$ total amount of probes.
3. Compute the rank sum W from this rank.
4. Compute the associated z value.
5. Compute the p -value as $p = P(Z \geq z)$ for a normal distribution $N(0, 1)$.
6. **If $p \leq 0.06$, assign a “detected” flag to the probes and probe set.**
7. Otherwise, assign a “not detected” flag to the probes and probe set. Normalization and Summarization Median Polish (See Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003 Apr.4(2):249-64.)

Additional Information

Q&A

1. Can you export data by single species selection instead of having all species included?
Answer: You cannot export by species. You can only sort by species later in another tool like Excel.
2. On the annotations we are not ‘compliant’ with the international standards, i.e., we use the word ‘star’ instead of ‘*’ for naming the probes that hybridize to the antisense mature mirna. Also mirna should be named miRNA is refers to a mature form. Why do our annotations not follow the standards?
Answer: The Windows OS and our Design software, do not allow “*” in probeset names or filenames, so we substituted in the expression “star”

Definitions

- **DAT file:** The image of the scanned probe array.
- **CEL file:** The software derives the *.CEL file from a *.DAT file and automatically creates it upon opening a *.DAT file. It contains a single intensity value for each probe cell delineated by the grid (calculated by the Cell Analysis algorithm).
- **CDF file:** x, y coordinates for probes, probe set name and probe sequences
- **QCC file:** QC file that instructs what control probe sets to incorporate
- **BGP:** we eliminated the canonical MM probe. This file lists antigenomic probes based on GC content

