# Agilent CytoGenomics 5.1

**Feature Extraction for CytoGenomics**

## Reference Guide

**For Research Use Only. Not for use in diagnostic procedures.**

**Agilent Technologies**

# Notices

## Manual Part Number

## Edition

## Patents

Portions of this product may be covered under US patent 6571005 licensed from the Regents of the University of California.

## Technical Support

### For US and Canada

Call 800-227-9770 (option 3, 4, 4)
Or send an email to:
informatics_support@agilent.com

### For Europe, Middle East, Africa, and India

Find the support center telephone number for your location at www.agilent.com/genomics under **Contact Us**. Or send an email to genomics_tech_europe@agilent.com

### For all other regions

Agilent's world-wide sales and support center telephone and email assistance information can be obtained at www.agilent.com/genomics under **Contact Us**.

## Warranty

## Technology Licenses

## Restricted Rights Legend

## Safety Notices

**CAUTION**

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

**WARNING**

**A WARNING notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a WARNING notice until the indicated conditions are fully understood and met.**

# In This Guide...

This *Reference Guide* contains tables that list default parameter values and results for Agilent Feature Extraction for CytoGenomics analyses, and explanations of how Feature Extraction for CytoGenomics uses its algorithms to calculate results.

## 1 Protocol Default Settings

This chapter includes tables that list the default parameter values found in the protocols shipped with the software.

## 2 QC Report Results

Learn how to read and interpret the QC Reports.

## 3 Text File Parameters and Results

This chapter contains a listing of parameters and results within the text file produced after Feature Extraction.

## 4 XML (MAGE-ML) Results

Refer to this chapter to find the results contained in the MAGE-ML files generated after Feature Extraction.

## 5 How Algorithms Calculate Results

Learn how Feature Extraction algorithms calculate the results that help you interpret your gene expression experiments.

## 6 Command Line Feature Extraction

This chapter contains the commands and arguments to integrate Feature Extraction into a completely automated workflow.

# Acknowledgments

# Content

# 1

# Default Protocol Settings

*See the Agilent Feature Extraction for CytoGenomics User Guide to learn the purpose of all the parameters and settings and how to modify them.*

*Agilent protocols are meant for use with Agilent microarrays and are intended for use with arrays that use Agilent default lab procedures (label, hybridization, wash, and scanning methods). The non-Agilent protocol is meant for use with non-Agilent microarrays that are scanned with an Agilent scanner.*

When a protocol is assigned to an extraction set, the software loads a set of protocol parameter values and settings that affect the process and results for Feature Extraction.

Parameter values in the protocol depend on the microarray type and your experiment. The following pages list the default settings for each of the protocol templates shipped or downloaded with the software. Each protocol template represents a different microarray type. You can display these settings and values when you open the Protocol Editor for each of the protocol templates.

**NOTE**

Note that Feature Extraction for CytoGenomics is not supported on Macintosh systems and is not installed with the Macintosh version of the Agilent CytoGenomics software. To run an analysis workflow, you must first extract the image file using the standalone Agilent Feature Extraction program (on a Windows PC), then use the extracted FE file to run a manual workflow in the Macintosh version of CytoGenomics.

# Default Protocol Settings – Introduction

*To learn more about changing the default values for the protocols, see the Agilent Feature Extraction for CytoGenomics User Guide.*

This chapter presents tables for display of the default settings for each protocol. Parameter values depend on:

- microarray type
- lab protocol
- formats
- scanner used

*To learn about the naming of the protocol templates, see the Agilent Feature Extraction for CytoGenomics User Guide.*

# Default Protocol Settings

**CAUTION**   These protocol settings may not be optimal for non-Agilent microarrays or Agilent microarrays processed with non-Agilent procedures. You must determine the settings and values that are optimal for your system.

## CytoCGH_0500_1x_Nov17
## CytoCGH_0500_2x_Nov17
## CytoCGH_0500_4x_Nov17
## CytoCGH_0500_8x_Nov17
## CytoCGH_0500_SingleCell_Nov17

These are CGH protocols for use with the *Oligonucleotide Array-Based CGH for Genomic DNA Analysis* (Enzymatic User Manual version 6.1 or higher, ULS User Manual version 3.1 or higher). The protocols come preloaded with Feature Extraction for CytoGenomics 5.1.

The *CytoCGH_0500_SingleCell_Nov17* protocol is for use with arrays of AMADID 067559 or 067649, which are designed for analysis of single cells. For all other arrays, the number of arrays per slide determines which protocol the program uses for extraction (the 1x protocol for single-pack format, the 2x protocol for 2-pack format, etc.).

**Table 1**    Default settings for the preloaded CGH protocols

| Protocol steps | Parameter | | Default Setting/Value (v10.10) |
|---|---|---|---|
| Place Grid | Array Format | For any format automatically determined or selected by you, the software uses the default Placement Method listed below. Parameters that apply to specific formats appear only if that format is selected. | Automatically Determine [Recognized formats: Single Density (11k, 22k), 25k, Double Density (44k), 95k, 185k, 185k 10 uM, 65 micron feature size (also with 10 micron scans), 30 micron feature size single pack and multi pack, and Third Party] |
| | | Placement Method | Hidden if Array Format is set to Automatically Determine. Allow Some Distortion (All formats) |
| | | Enable Background Peak Shifting | Hidden if Array Format is set to Automatically Determine. Set to False for all arrays except 30 microns single pack and multi pack, for which it is set to True. |
| | | Use central part of pack for slope and skew calculation? | Hidden if Array Format is set to Automatically Determine. Set to False for all arrays except 30 microns single pack and multi pack, for which it is set to True. |
| | | Use the correlation method to obtain origin X of subgrids | Hidden if Array Format is set to Automatically Determine. Set to False for all arrays except 30 microns single pack and multi pack, for which it is set to True. |

**Table 1** Default settings for the preloaded CGH protocols (continued)

| Protocol steps | Parameter | | Default Setting/Value (v10.10) |
|---|---|---|---|
| | Use Enhanced Gridding | Apply the enhanced gridding feature released in Feature Extraction for CytoGenomics version 5.0. The enhancements include a new iterative method for determining grid position, rotation, and skew, and several "fine" grid tuning methods that improve the calculation of rotation and skew. Enhanced gridding also uses both the foreground and background of the corner stencil patterns to improve identification of grid corners. | True |
| Optimize Grid Fit | Grid Format | The parameters and values for optimizing the grid differ depending on the format. | Automatically Determine [Recognized formats: 65 micron feature size, 30 micron feature size, and Third Party] |
| | | Iteratively Adjust Corners? | Hidden if Array Format is set to Automatically Determine. True (All Formats, except Third Party) False (Third Party) |
| | | Adjust Threshold | Hidden if Array Format is set to Automatically Determine. 0.300 (All Formats, except Third Party) |
| | | Maximum Number of Iterations | Hidden if Array Format is set to Automatically Determine. 5 (All Formats, except Third Party) |
| | | Found Spot Threshold | Hidden if Array Format is set to Automatically Determine. 0.200 (All Formats, except Third Party) |
| | | Number of Corner Feature Side Dimension? | Hidden if Array Format is set to Automatically Determine. 20 (All Formats, except Third Party) |

**Table 1** Default settings for the preloaded CGH protocols (continued)

| Protocol steps | Parameter | | Default Setting/Value (v10.10) |
|---|---|---|---|
| Find Spots | Spot Format | Depending on the format selected by the software or by you, the default settings for this step change. See the rows below for the default values for finding spots. | Automatically Determine [Recognized formats: Single Density (11k, 22k), 25k, Double Density (44k), 95k, 185k, 185k 10 uM, 244k 10uM, 65 micron feature size, 30 micron feature size, and Third Party] |
| | | Use the Nominal Diameter from the Grid Template | Hidden if Array Format is set to Automatically Determine. True (All Formats) |
| | | Spot Deviation Limit | Hidden if Array Format is set to Automatically Determine. 8.0 for all formats except for third-party, for which it is set to 1.5 |
| | | Calculation of Spot Statistics Method | Hidden if Array Format is set to Automatically Determine. Use Cookie (All Formats) |
| | | Cookie Percentage | Hidden if Array Format is set to Automatically Determine. 0.650 (Single Density, 25k) |
| | | | 0.561 (Double Density, 95k) |
| | | | 0.700 (185k, 185k 10 uM, 244k 10 uM, 65 micron feature size) |
| | | | 0.750 (30 micron feature size) |
| | | Exclusion Zone Percentage | Hidden if Array Format is set to Automatically Determine. 1.200 (All Formats except 30 micron feature size) |
| | | | 1.300 (30 micron feature size) |
| | | Auto Estimate the Local Radius | Hidden if Array Format is set to Automatically Determine. True (Single Density, Double Density, 25k, 95k) |

**Table 1**    Default settings for the preloaded CGH protocols (continued)

| Protocol steps | Parameter | | Default Setting/Value (v10.10) |
|---|---|---|---|
| | | | False (185k, 185k 10uM, 65 micron feature size, 30 micron feature size, 244k 10uM) |
| | LocalBGRadius | | Hidden if Array Format is set to Automatically Determine. 100 (when False for 185k, 185k 10uM, 65 micron feature size, 244k 10 uM) |
| | | | 150 (when False for 30 micron feature size) |
| | Pixel Outlier Rejection Method | | Inter Quartile Region (Automatically Determine and All Formats) |
| | | RejectIQRFeat | 1.42 (All Formats) |
| | | RejectIQRBG | 1.42 (All Formats) |
| | Statistical Method for Spot Values from Pixels | | Use Mean/Standard Deviation (Automatically Determine and All Formats) |
| Flag Outliers | Compute Population Outliers | | True |
| | | Minimum Population | 10 |
| | | IQRatio | 1.42 |
| | | Background IQRatio | 1.42 |
| | | Use Qtest for Small Populations? | True |
| | | Report Population Outliers as Failed in MAGEML file | False |
| | Compute Non Uniform Outliers | | True |
| | Scanner | The values for the parameters change depending on the scanner used for the image. See below for differences. | Automatically Determine |
| | **Agilent scanner** | | |

**Table 1** Default settings for the preloaded CGH protocols (continued)

| Protocol steps | Parameter | | Default Setting/Value (v10.10) |
|---|---|---|---|
| | Automatically Compute OL Polynomial Terms | | Hidden if Array Format is set to Automatically Determine. True |
| | | Feature – (%CV)^2 | 0.04000 |
| | | Red Poissonian Noise Term Multiplier | 5 |
| | | Red Signal Constant Term Multiplier | 1 |
| | | Green Poissonian Noise Term Multiplier | 5 |
| | | Green Signal Constant Term Multiplier | 1 |
| | | Background – (%CV)^2 | 0.09000 |
| | | Red Poissonian Noise Term Multiplier | 3 |
| | | Red Background Constant Term Multiplier | 1 |
| | | Green Poissonian Noise Term Multiplier | 3 |
| | | Green Background Constant Term Multiplier | 1 |
| Compute Bkgd, Bias and Error | Background Subtraction Method | | No Background Subtraction |
| | Significance (for IsPosAndSignif and IsWellAboveBG) | | Use Error Model for Significance |
| | | 2-sided t-test of feature vs. background max p-value | 0.01 |
| | | WellAboveMulti | 13 |
| | Signal Correction—Calculate Surface Fit (required for Spatial Detrend) | | True |
| | | Feature Set for Surface Fit | OnlyNegativeControlFeatures |
| | | Perform Filtering for Surface Fit | False |

**Table 1** Default settings for the preloaded CGH protocols (continued)

| Protocol steps | Parameter | | Default Setting/Value (v10.10) |
|---|---|---|---|
| | | Perform Spatial Detrending | True |
| | Signal Correction—Adjust Background Globally | | False |
| | Signal Correction—Perform Multiplicative Detrending | | True |
| | | Detrend on Replicates Only | False |
| | | Filter Low signal probes from Fit? | True |
| | | Neg. Ctrl. Threshold Mult. Detrend Factor | 3 |
| | | Perform Filtering for Fit | Use Window Average |
| | | Use polynomial data fit instead of LOESS? | True |
| | | Polynomial Multiplicative DetrendDegree | 4 |
| | Robust Neg Ctrl Stats? | | True |
| | Choose universal error, or most conservative | | Most Conservative |
| | | MultErrorGreen | 0.1000 |
| | | MultErrorRed | 0.1000 |
| | | Auto Estimate Add Error Red | True |
| | | Auto Estimate Add Error Green | True |
| | Use Surrogates | | True |
| Correct Dye Biases | Use Dye Norm List | | Automatically Determine |
| | Dye Normalization Probe Selection Method | | Use Rank Consistent Probes |
| | | Rank Tolerance | 0.050 |
| | | Variable Rank Tolerance | False |
| | Signal Characteristics | | OnlyPositiveAndSignificantSignals |
| | Normalization Correction Method | | Linear |
| | | Max Number Ranked Probes | -1 |

**Table 1** Default settings for the preloaded CGH protocols (continued)

| Protocol steps | Parameter | Default Setting/Value (v10.10) |
|---|---|---|
| | Omit Background Population Outliers | False |
| | Allow Positive and Negative Controls | False |
| Compute Ratios | Peg Log Ratio Value | 4.00 |
| Calculate Metrics | Grid Test Format | Automatically Determine Recognized formats: 60 and 30 micron feature size, third-party |
| | Spikein Target Used | False |
| | Min Population for Replicate Stats? | 3 |
| | PValue for Differential Expression | 0.010000 |
| | Percentile Value | 75.00 |
| Generate Results | Type of QC Report | Streamlined CGH |
| | Generate Single Text File | True |
| | JPEG Down Sample Factor | 4 |

# 2
# QC Report Results

QC reports include statistical results to help you evaluate the reproducibility and reliability of your single microarray data. Use plots and statistics from the report to:

- Set up your own run charts of statistical values versus time or experiment number to track performance of one microarray compared to other microarrays

- Monitor upstream lab protocols, such as performance of your hybridization/washing steps

- Monitor the effect of changing Feature Extraction protocol parameters on the performance of your data analysis

If you incorporate a set of QC metrics in your extraction, those results will appear on the final page of the QC report as an Evaluation Table.

**Agilent Technologies**

# QC Reports

## Streamlined CGH QC Report

The streamlined CGH QC report provides QC metrics that are relevant to CGH application. All log plots use log base 2 (not 10).

**1** *"QC Report Headers" on page 27*

**2** *"Spot Finding of Four Corners" on page 28*

**3** *"Spatial Distribution of All Outliers" on page 29*

**4** *"QC reports with metric sets added" on page 24*

**5** *"Histogram of Signals Plot" on page 34*

**6** *"Outlier Stats" on page 29*
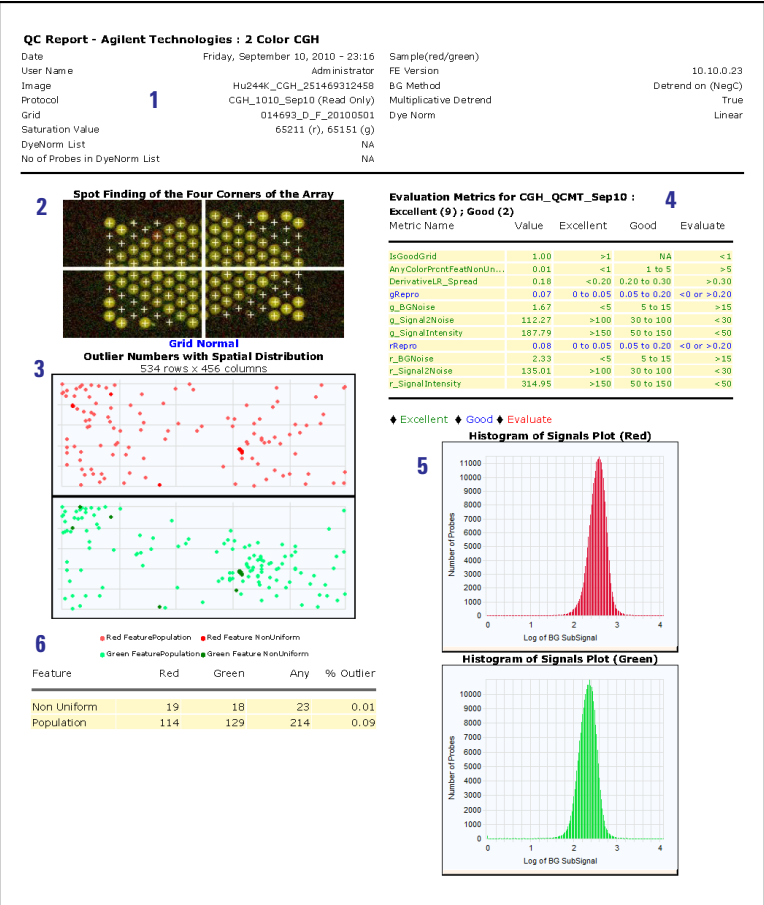


**Figure 1**     Streamlined CGH QC Report (p1)

22     **Feature Extraction for CytoGenomics 5.1 Reference Guide**

**7** "**Spatial Distribution of Significantly Up-Regulated and Down-Regulated Features (Positive and Negative Log Ratios)**" **on page 37**

**8** "**Plot of Background-Corrected Signals**" **on page 33**
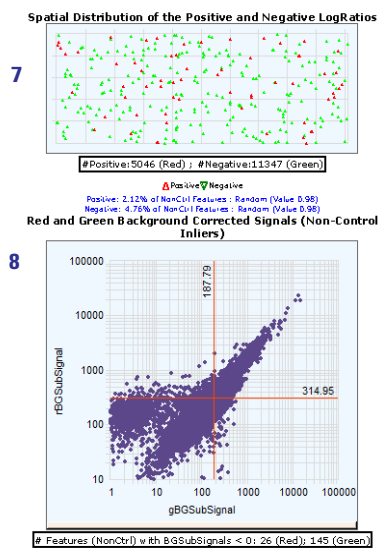


**Figure 2**     Streamlined CGH QC Report (p2)

# QC reports with metric sets added

When metric sets are associated to the protocols, QC reports are generated with an additional set of evaluation metrics. Depending on the microarray types, some QC metric sets come with thresholds (denoted by QCMT) and some without thresholds (denoted by QCM).

If thresholds are included in the metric set, the evaluation tables in the QC report show metrics that are within threshold ranges or that have exceeded those ranges.

Agilent has determined which of the FE Stats are good metrics to follow the processing of our arrays. Most of the metrics chosen will be useful to determine if there are problems in the various laboratory steps (label, hybridization, wash, scan steps). The new "IsGoodGrid" metric tracks the automatic grid-finding of Feature Extraction. By looking at a lot of data run on our arrays, using our wet-lab protocols, Agilent has found thresholds that indicate if the data is in the expected range ("Good") or out of the expected range ("Evaluate").

For some applications (CGH, miRNA), an extra threshold level, "Excellent" is provided. More data has been screened to allow us to set the metric thresholds to a tighter limit that indicate excellent processing. For those applications that do not have a full set of thresholds (e.g. ChIP), or no "Excellent" thresholds (e.g. GE1 and GE2), the user should be assured that the data coming from the "Good" grade is good to use. Excellent thresholds for those applications may be provided in the future.

## QC metric set results--default protocol settings

Figure 3 is an example of part of a QC report — the header and the Evaluation Metrics table — generated from a 2-color gene expression extraction whose GE2 metric set with thresholds had been added. In this extraction the default protocol settings were used. Note that all values for the metrics are within the default threshold ranges.

**QC Report - Agilent Technologies : 2 Color Gene Expression**

| | | | |
|---|---|---|---|
| Date | Saturday, September 11, 2010 - 01:20 | BG Method | No Background |
| Image | Hu22K_GE2_251209710036 | Background Detrend | On(FeatNCRange, LoPass) |
| Protocol | GE2_1010_Sep10 (Read Only) | Multiplicative Detrend | True |
| User Name | Administrator | Dye Norm | Linear Lowess |
| Grid | 012097_D_20070820 | Linear DyeNorm Factor | 4.15(Red) 15.9(Green) |
| FE Version | 10.10.0.23 | Additive Error | 14(Red)65(Green) |
| Sample(red/green) | | Saturation Value | 65211 (r), 65185 (g) |
| DyeNorm List | NA | | |
| No of Probes in DyeNorm List | NA | | |

**Evaluation Metrics for GE2_QCMT_Sep10 :**
**Good (12)**

| Metric Name | Value | Excellent | Good | Evaluate |
|---|---|---|---|---|
| IsGoodGrid | 1.00 | | >1 | <1 |
| AnyColorPrcntFeatNonUn… | 0.05 | | <1 | >1 |
| gNegCtrlAveBGSubSig | -3.61 | | -20 to 10 | <-20 or >10 |
| gNegCtrlSDevBGSubSig | 2.83 | | <15 | >15 |
| rNegCtrlAveBGSubSig | -3.93 | | -20 to 4 | <-20 or >4 |
| rNegCtrlSDevBGSubSig | 2.07 | | <6 | >6 |
| gNonCntrlMedCVBkSubSig… | 13.48 | | 0 to 18 | <0 or >18 |
| rNonCntrlMedCVBkSubSig… | 15.05 | | 0 to 18 | <0 or >18 |
| gE1aMedCVBkSubSignal | 10.57 | | 0 to 18 | <0 or >18 |
| rE1aMedCVBkSubSignal | 10.21 | | 0 to 18 | <0 or >18 |
| absE1aObsVsExpCorr | 0.99 | | >0.86 | <0.86 |
| absE1aObsVsExpSlope | 0.95 | | >0.85 | <0.85 |

♦ Excellent  ♦ Good  ♦ Evaluate

**Figure 3**  Partial QC Report—Header and Evaluation Metrics with GE2 metric set with thresholds added—Default protocol settings

## QC metric set results—Spatial and Multiplicative Detrending Off

Figure 4 is an example of a QC report header and Evaluation Metrics table generated from a 2-color gene expression extraction whose GE2 metric set with thresholds had been added. In this extraction spatial and multiplicative detrending were turned off. Note that not all values of the metrics are within the default thresholds.

**QC Report - Agilent Technologies : 2 Color Gene Expression**

| | | | |
|---|---|---|---|
| Date | Saturday, September 11, 2010 - 01:42 | BG Method | No Background |
| Image | Hu22K_GE2_251209710036 | Background Detrend | Off |
| Protocol | GE2_1010_Sep10_2_2 (Editable) | Multiplicative Detrend | False |
| User Name | Administrator | Dye Norm | Linear Lowess |
| Grid | 012097_D_20070820 | Linear DyeNorm Factor | 4.05(Red) 6.84(Green) |
| FE Version | 10.10.0.23 | Additive Error | 13(Red)28(Green) |
| Sample(red/green) | | Saturation Value | 65211 (r), 65185 (g) |
| DyeNorm List | NA | | |
| No of Probes in DyeNorm List | NA | | |

**Evaluation Metrics for GE2_QCMT_Sep10 :**
**Good (10) ; Evaluate (2)**

| Metric Name | Value | Excellent | Good | Evaluate |
|---|---|---|---|---|
| IsGoodGrid | 1.00 | | >1 | <1 |
| AnyColorPrcntFeatNonUn... | 0.05 | | <1 | >1 |
| gNegCtrlAveBGSubSig | 77.46 | | -20 to 10 | <-20 or >10 |
| gNegCtrlSDevBGSubSig | 3.59 | | <15 | >15 |
| rNegCtrlAveBGSubSig | 66.53 | | -20 to 4 | <-20 or >4 |
| rNegCtrlSDevBGSubSig | 2.19 | | <6 | >6 |
| gNonCntrlMedCVBkSubSig... | 4.78 | | 0 to 18 | <0 or >18 |
| rNonCntrlMedCVBkSubSig... | 6.56 | | 0 to 18 | <0 or >18 |
| gE1aMedCVBkSubSignal | 7.64 | | 0 to 18 | <0 or >18 |
| rE1aMedCVBkSubSignal | 9.27 | | 0 to 18 | <0 or >18 |
| absE1aObsVsExpCorr | 0.97 | | >0.86 | <0.86 |
| absE1aObsVsExpSlope | 0.87 | | >0.85 | <0.85 |

♦ Excellent  ♦ Good  ♦ Evaluate

**Figure 4**    QC Report Header and Evaluation Metrics with GE2 metric set with thresholds added—Detrending turned off

# QC Report Headers

## Streamlined CGH QC Report

The streamlined CGH QC report contains the same header information as the 2‑color gene expression QC report, except for Linear DyeNorm Factor and Additive Error which are removed. Also, the information from the two fields, "BG Method" and "Background Detrend", have been collapsed into the one field, "BG Method".

## CGH_ChIP QC Report

All header information that appears in the 2‑color gene expression QC report are included in the CGH_ChIP report. This report lists one additional metric, **Derivative of Log Ratio Spread** in the header information.

**Derivative of Log Ratio Spread**  Measures the standard deviation of the probe‑to‑probe difference of the log ratios. This is a metric used in CGH experiments where differences in the log ratios are small on average. A smaller standard deviation here indicates less noise in the biological signals.

# Feature Statistics

This section provides an explanation for each of the feature statistics segments of the QC report and how these feature statistics can help you assess the performance of your microarray system.

## Spot Finding of Four Corners

By looking at the features in the four corners of the microarray, you can decide if the spot centroids have been located properly. If their locations are off-center in one or more corners, you may have to run the extraction again with a new grid.
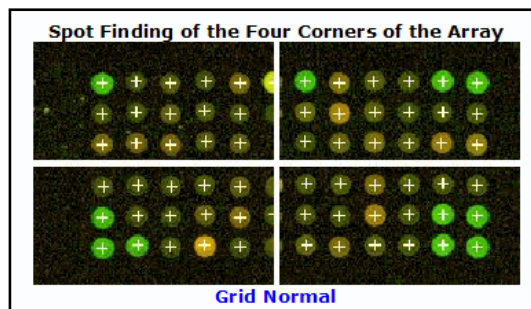


**Figure 5**     QC Report—Spot Finding for Four Corners

# Outlier Stats

If the QC Report shows a greater than expected number of non-uniform or population outliers, you may want to check your hybridization/wash step. Also, check the visual results (.shp file) to see if the spot centroids are off-center. If the grid was not placed correctly, a new grid is required.

| | Feature | | Local Background | |
|---|---|---|---|---|
| | Red | Green | Red | Green |
| Non Uniform | 4 | 8 | 9 | 0 |
| Population | 98 | 73 | 48 | 0 |

**Figure 6**     QC Report—Outlier Stats

For 1-color reports, the number of outliers is reported for the green channel only.

# Spatial Distribution of All Outliers

The QC report shows two plots of all the outliers, both population and nonuniformity outliers, whose positions are distributed across the microarray. One plot is for the green channel, and the other, for the red channel. SNP probes are included.

To distinguish the background population and nonuniform outliers from one another, look at the color coding at the bottom of the two plots.

For the 1-color report, only the green plot is shown.
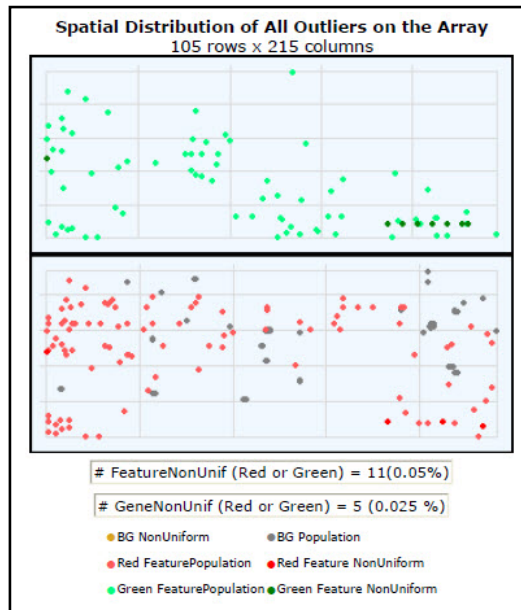
**Figure 7**     QC Report—Number and Spatial Distribution of Outliers

The number (and percentage) of features that are feature nonuniformity outliers in either the green or red channel is shown below the plot. The 1-color report shows only the percentage of green feature non-uniformity outliers.

Also, the number (and percentage) of genes that are nonuniformity outliers in either channel is shown below the plot. If there were replicate features representing one gene and at least one feature was not an outlier, no gene outliers would appear.

# Net Signal Statistics

*Net signal* is the mean signal minus the scanner offset. Net signal is used so that these statistics are independent of the scanner version.

Net signal statistics are an indication of the dynamic range of the signal on a microarray for both non-control probes and spike-in probes (not applicable for CGH QC report). The QC Report uses the range from the 1st percentile to the 99th percentile as an indicator of dynamic range for that microarray. NetSignal is also a column in the FeatureData output.

For example, in the figure below for non-control probes the dynamic range of the net signal intensity for the red channel is from 42 to 6803 with half the probes having a net signal intensity of greater than the median of 97 and half below the median of 97. The median (or 50th percentile) represents the middle of the ranked-values of the distribution of signals.

Another indicator of signal range for the microarray is the number of features that are saturated in the scanned image (i.e., NumSat).

**Net Signal Statistics**

**Agilent SpikeIns:**

|  | Red | Green |
|---|---|---|
| # Saturated Features | 0 | 0 |
| 99% of Sig. Distrib. | 24937 | 11899 |
| 50% of Sig. Distrib. | 2351 | 750 |
| 1% of Sig. Distrib. | 160 | 112 |

**Non-Control probes:**

|  | Red | Green |
|---|---|---|
| # Saturated Features | 16 | 1 |
| 99% of Sig. Distrib. | 6850 | 1750 |
| 50% of Sig. Distrib. | 82 | 64 |
| 1% of Sig. Distrib. | 40 | 48 |

**Figure 8**     QC Report—Net Signal Statistics

# Negative Control Stats

The Negative Control Stats table includes the average and standard deviation of the net signals (mean signal minus scanner offset) and the background-subtracted signals for both the red and green channels in the negative controls. These statistics filter out saturated and feature non-uniform and population outliers and give a rough estimate of the background noise on the microarray. SNP probes are not included in these statistics.

| Negative Control Stats | Red | Green |
|---|---|---|
| Average Net Signals | 33.82 | 19.60 |
| StdDev Net Signals | 5.85 | 5.27 |
| Average BG Sub Signal | 0.37 | 0.33 |
| StdDev BG Sub Signal (BG Noise) | 5.64 | 4.96 |

**Figure 9** QC Report—Negative Control Stats

# Plot of Background-Corrected Signals

Figure 10 is a plot of the log of the red background-corrected signal versus the log of the green background-corrected signal for non-control inlier features. The linearity or curvature of this plot can indicate the appropriateness of background method choices. The plot should be linear.

The intersection of the red vertical and horizontal lines shows the location of the median signal. The numbers along the edge of the lines represent the location of the median signal on the plot.

The values below the plot indicate the number of non-control features that have a background-corrected signal less than zero. SNP probes are not included.
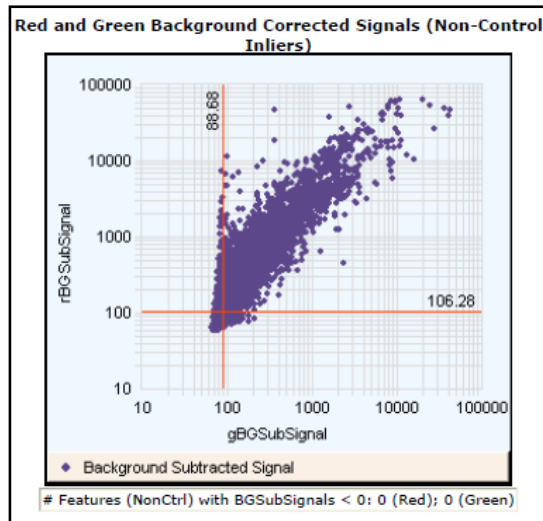


**Figure 10**    QC Report—Plot of Background-Corrected Signals

# Histogram of Signals Plot

The purpose of this histogram is to show the level of signal and the shape of the signal distribution. The histogram is a line plot of the number of points in the intensity bins vs. the log of the processed signal. SNP probes are not included.
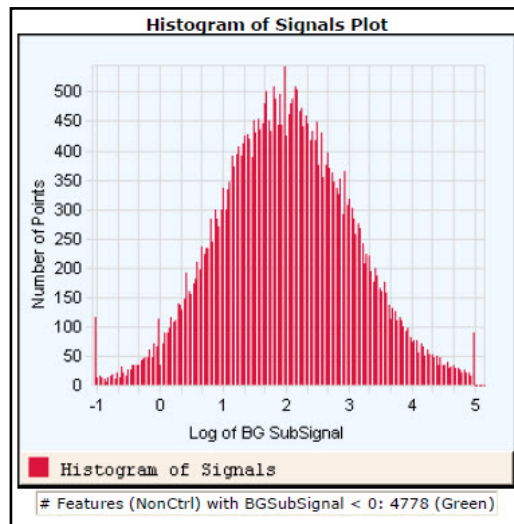


**Figure 11**    1-color QC Report—Histogram of Signals Plot

# Local Background Inliers

With these numbers you can see the mean signal distribution for the local background regions (BGMeanSignal) after outliers have been removed. This information can help you detect hybridization/wash artifacts and can be a component of noise in the low signal range. SNP probes are included.

| Local Bkg (inliers) | | |
|---|---|---|
| | Red | Green |
| Number | 22105 | 22153 |
| Avg | 49.77 | 41.00 |
| SD | 0.93 | 1.35 |

**Figure 12**    QC Report—Local Background Inliers

# Foreground Surface Fit

Spatial Detrend attempts to account for low signal background that is present on the feature "foreground" and varies across the microarray. SNP probes are not included.

- A high RMS_Fit number can indicate gradients in the low signal range before detrending.

- RMS_Resid indicates residual noise after detrending.

- AvgFit indicates how much signal is in the "foreground".

  A higher AvgFit number indicates a larger amount of signal was detected by the detrend algorithm and removed.

  This value may include the scanner offset, unless a background method has been used before detrending. The value may not include higher frequency background signals. These higher frequency background signals are best removed by using the Local Background Method before the detrending algorithm.

**Figure 13**   QC Report—Foreground Surface Fit

# Multiplicative Surface Fit

This is the root mean square (RMS) of the surface fit for the data. The RMS X 100 is roughly the average % deviation from "flat" on the microarray. A multiplicative trend means that there are regions of the microarray that are brighter or dimmer than other regions. This trend is an effect that multiplies signals; that is, a brighter signal is more affected in absolute signal counts than a dimmer signal. SNP probes are not included in calculation of multiplicative detrending.

If the signal is improved through a multiplicative surface fit, the RMS_Fit value appears as a fraction, as in the figure below.



**Figure 14**   QC Report—Multiplicative Surface Fit

**What if multiplicative detrending does not work?**

If the median %CV for the Processed Signal of the non-control probes is greater than the BGSub Signal median %CV after multiplicative detrending, Feature Extraction turns off multiplicative detrending.

The QC report shows an RMS_Fit = 0.0 if multiplicative detrending did not result in better data.

If there are no stats for non-control probes, Feature Extraction looks at the spike-in control probes. If the %CVs for these become worse, Feature Extraction removes detrending.

If the option "Detrend on Replicates only" is chosen and if there are not enough replicates for non-control or spike-in control probes, Feature Extraction turns off multiplicative detrending.

# Spatial Distribution of Significantly Up-Regulated and Down-Regulated Features (Positive and Negative Log Ratios)

You can display the distribution of the significantly up- and down-regulated features on this plot (up–red; down–green).
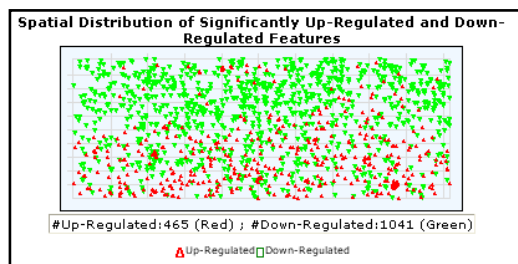


**Figure 15**   QC Report—Spatial Distribution of Up- and Down-Regulated Features

For the CGH QC Report, this plot is referred to as "Spatial Distribution of the Positive and Negative Log Ratios".

If the microarray contains greater than 5000 features, the software randomly selects 5000 data points. These points include the number of up-regulated features in the same proportion to the number of down-regulated features as they are found on the actual microarray.

The threshold that is used to determine significance is set in the protocol—QCMetrics_differentialExpressionPValue.

These are the same features shown as up- or
down-regulated in Figure 16.

## Plot of LogRatio vs. Log ProcessedSignal

This plot shows the log ratios of non-control inliers vs. the
log of their red and green processed signals. The color
coding signifies the degree to which features are significantly
differentially expressed: those that are up-regulated (red),
those that are down-regulated (green) and those that cannot
confidently be said to show gene expression (light yellow).

For the CGH QC Report, these are referred to as "Positive",
"Negative" log ratios (base 2). The threshold that is used to
determine significance is set in the protocol
(QCMetrics_differentialExpressionPValue).

Features that were used for normalization are indicated in
blue. Significance takes precedence over normalization for
the color coding; that is, features that are both significantly
differentially expressed and used for normalization will be
color-coded either red or green. SNP probes are not
included.

*LogProcessedSignal in the plot is [Log(rProcessedSignal x gProcessedSignal)]/2.*
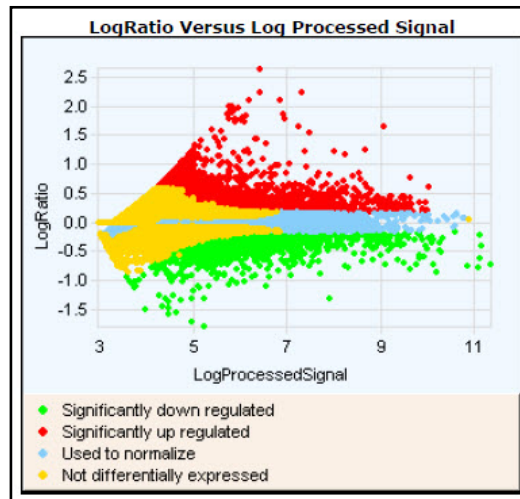


**Figure 16**    QC Report—Plot of Up- and Down-Regulated Features

## Spatial Distribution of Median Signals for each Row and Column

The first of these graphs plots the median Processed Signal and median BGSub Signal for each row over all columns of a 1-color GE microarray. The second plots the same signals for each column over all rows of the 1-color GE microarray. The difference between the Processed Signal and the BGSubSignal represents the effect of the multiplicative detrending. The Processed Signal should look flatter.

*Higher frequency noise is shown in these plots so you can distinguish a low frequency trend outside of the high frequency noise.*
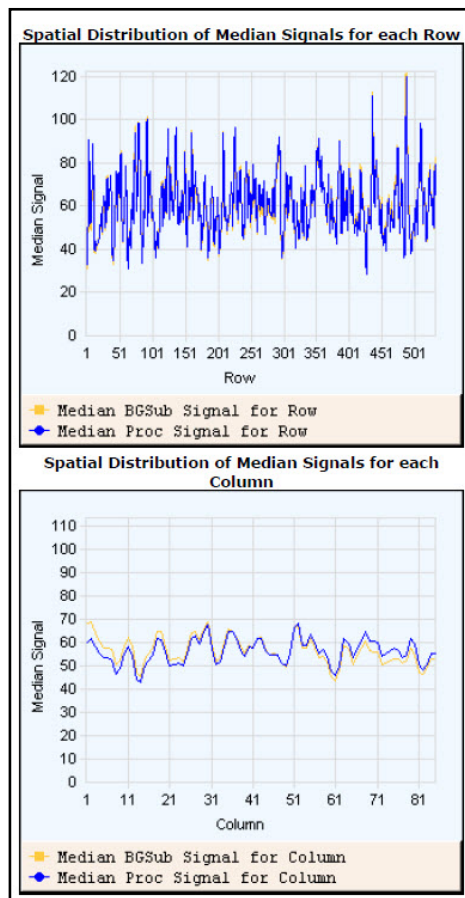


**Figure 17**    1-color QC Report—Median Signal Spatial Distribution

## Histogram of LogRatio plot

This is a plot of the log ratio distributions, and displays the log ratios vs. the number of probes. This plot is included only in the CGH_ChIP report, which is the default report for the ChIP_1010_Sep10 protocol.
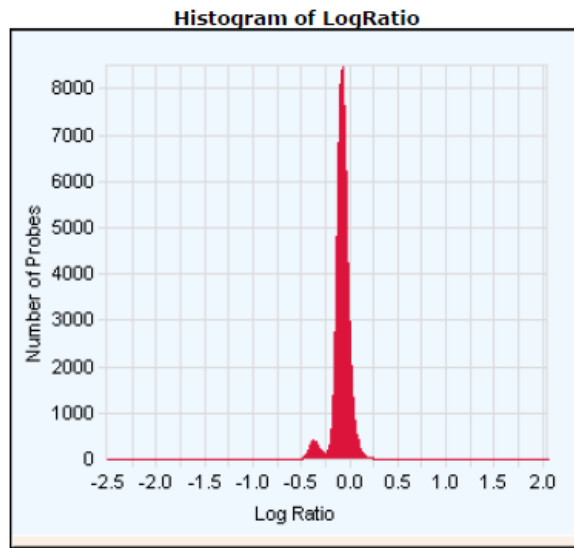
**Figure 18**  Histogram of LogRatio plot

# Inter-Feature Statistics

*Spike-in probes* are known probes that are hybridized with known quantities of a target "spike-in" cocktail. They are used to perform a quality check of the microarray/experiment.

Some microarray designs have replicated non-control probes; that is, multiple features on the microarray contain the same probe sequence. Many of the Agilent microarray designs also have *spike-in probes*, which are replicated across the microarray (e.g., some microarrays have 10 sequences with 30 replicates each). The QC Report uses these replicated probes to evaluate reproducibility of both the signals and the log ratios. Metrics such as signal %CV and log ratio statistics are calculated if probes are present with a minimum number of replicates.

The protocol indicates if labeled target to these spike-in probes has been added in the hybridization (QCMetrics_UseSpikeIns). The minimum number of replicates (inliers to Sat & NonUnif flagging) is also set in the protocol (QCMetrics_minReplicate Population).

This section provides an explanation for each of the segments of the QC report that cover inter-feature statistics and how these replicate statistics can help you assess performance.

## Reproducibility Statistics (%CV Replicated Probes)

### Non-control probes

If a non-control probe has a minimum number of inliers, a %CV (percent coefficient of variation) of the background-corrected signal is calculated for each channel (SD of signals/average of signals). This calculation is done for each replicated probe, and the median of those %CV's is reported in the table for each channel. SNP probes are not included.

| Reproducibility: %CV for Replicated Probes | | | | |
|---|---|---|---|---|
| | Median %CV Signal (inliers) | | | |
| | Non-Control probes | | Agilent SpikeIns | |
| | Red | Green | Red | Green |
| BGSubSignal | 15.05 | 13.48 | 10.21 | 10.57 |
| ProcessedSignal | 7.39 | 7.75 | 4.44 | 5.54 |

**Figure 19** QC Report—Reproducibility

A lower median %CV value indicates better reproducibility of signal across the microarray than a higher value.

## Exclusion of dim probes

Feature Extraction calculates the Median %CV using those probes bright enough to be in the range where the noise is more proportional to signal. Feature Extraction excludes from the calculation any sequences for which the Average (BGSubSignal) x Multiplicative error < Additive error/Dye Norm Factor. For 1-color data the Dye Norm Factor is 1.

A probe sequence will have a %CV calculated if the number of features that pass the filters (NonUniform and signal filter, described above) is greater than the minimum replicate number indicated in the protocol: "QCMetrics_minReplicatePopulation".

If the number of replicated sequences with enough inlier features is less than 10 or less than 10% of the replicated sequence, that is, if there are not enough bright replicated probes, the Median %CV field shows up as -1.

## Spike-in probes

The same algorithm is used to calculate the Median %CV for the spike-in probes as well. Because there are only 10 sequences in total and some are expected to fail the Additive error test described above, the minimum number of "bright enough" sequences required to calculate the Median %CV is 3.

# Microarray Uniformity (2-color only)

The QC Report has two metrics that measure the uniformity of replicated log ratios and that indicate the span of log ratios: average S/N and AbsAvgLogRatio. These are calculated from inlier features of replicated non-control and spike-in probes.

For example, some microarrays have 100 different non-control probe sequences with 10 replicate features each. For each replicate probe, the average and SD of the log ratios are calculated. The signal to noise (S/N) of the log ratio for each probe is calculated as the absolute of the average of the log ratios divided by the SD of the log ratios. From the population of 100 S/N's, for example, the average S/N is determined and shown in the table below.

The second metric, AbsAvgLogRatio, indicates the amount of differential expression (up-regulated or down-regulated). As described above, averages of log ratios are calculated for each replicated probe. The absolute of these averages is determined next. Then, the average of these absolute of averages is calculated to get a single value for the QC Report. The larger this value, the more differential expression is present.

| Array Uniformity: LogRatios | Non-Control | Agilent SpikeIns |
|---|---|---|
| AbsAvgLogRatio | 0.26 | 0.48 |
| AverageS/N | 3.86 | 43.07 |

**Figure 20**  QC Report—Array Uniformity: LogRatios

# Sensitivity

These values represent the NetSignal to background (BGUsed - ScannerOffset) ratio of the two spike-in probes with the lowest background-subtracted signal. Their purpose is to characterize the sensitivity of detecting a low signal relative to the background.

| Sensitivity:Agilent SpikeIns - Ratio of Signal to Background for 2 dimmest probes | | | |
|---|---|---|---|
| (+)E1A_r60_n11 | | (+)E1A_r60_a97 | |
| (g) | (r) | (g) | (r) |
| 4.1 | 3.2 | 15.0 | 2.3 |

**Figure 21**  QC Report—Sensitivity: Agilent SpikeIns Ratio of Signal to Background for 2 dimmest probes

# Reproducibility Plots

### Reproducibility plot for 2-color gene expression (spike-in probes)

Signal replicate statistics are calculated for spike-in probes if three criteria are met:

- They are present on the microarray.
- The protocol indicates that labeled target to these spike-in probes has been added in the hybridization (QCMetrics_UseSpikeIns is True).
- There are a minimum number of inlier features for calculations (QCMetrics_minReplicatePopulation).

As described above for non-control probes, %CV's are calculated for inliers for both red and green background-corrected signals. The %CV for each probe is plotted on the next page vs. the average of its background-corrected signal. The median of these %CV's is shown directly beneath the plot.
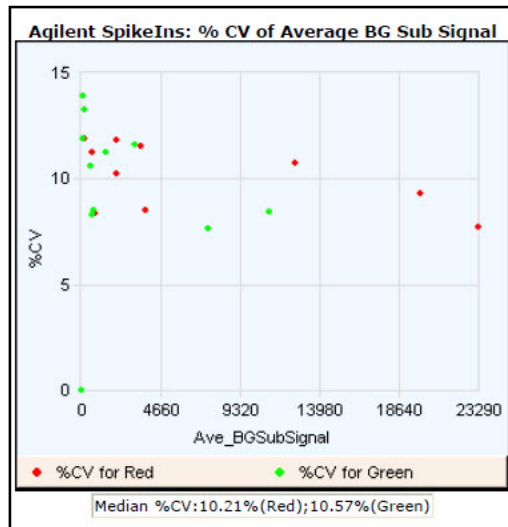
**Figure 22**  QC Report—Agilent SpikeIns: %CV of Average BGSub Signal

**Reproducibility plot for 1-color gene expression (spike-in probes)**

This graph plots %CV vs. the log_gMedianProcessedSignal for the 1-color gene expression microarray experiment. The region where the %CV flattens out and is not tightly correlated with signal is the range where noise is proportional to signal. This is generally the range used to calculate the median %CV.
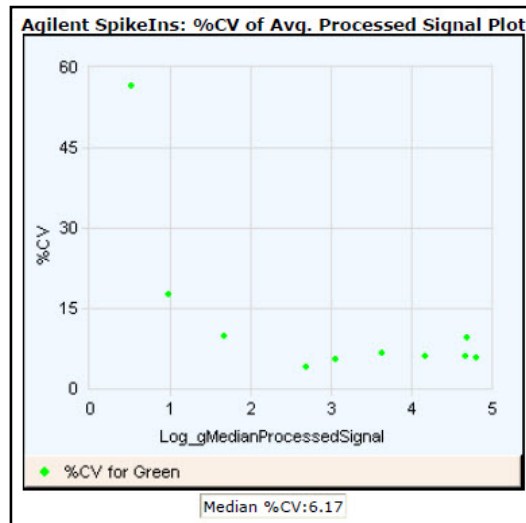
**Figure 23** 1-color QC Report—Agilent SpikeIns: %CV of Avg. Processed Signal Plot

### Reproducibility plot for miRNA (non-control probes)

This graph plots %CV vs. the log_gMedianProcessedSignal for the 1-color miRNA microarray experiment. The region where the %CV flattens out and is not tightly correlated with signal is the range where noise is proportional to signal. This is generally the range used to calculate the median %CV.
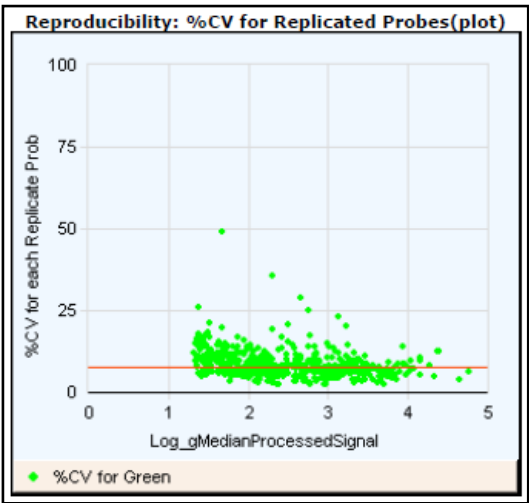


**Figure 24**    miRNA QC Report — Reproducibility: % CV for Replicated Probes

# Spike-in Signal Statistics

### 2-color gene expression spike-in signal statistics

These signal statistics and S/N values for spike-ins indicate accuracy and reproducibility of the signals of the microarray probes. The table shows the expected signal of the spike-in probe, the observed average signal, the SD of the observed signal and the S/N of the observed signal.

| Agilent SpikeIns Signal Statistics | | | | |
|---|---|---|---|---|
| Probe Name | Exp | Obs | SD | S/N |
| (+)E1A_r60_n9 | -1.00 | -1.10 | 0.01 | 106.67 |
| (+)E1A_r60_a107 | -0.48 | -0.53 | 0.01 | 56.44 |
| (+)E1A_r60_a135 | -0.48 | -0.47 | 0.01 | 53.69 |
| (+)E1A_r60_n11 | -0.48 | -0.56 | 0.02 | 32.96 |
| (+)E1A_r60_1 | 0.00 | 0.04 | 0.01 | 4.48 |
| (+)E1A_r60_a20 | 0.00 | -0.19 | 0.01 | 19.76 |
| (+)E1A_r60_3 | 0.48 | 0.44 | 0.02 | 28.48 |
| (+)E1A_r60_a104 | 0.48 | 0.35 | 0.01 | 34.03 |
| (+)E1A_r60_a97 | 0.48 | 0.34 | 0.02 | 14.13 |
| (+)E1A_r60_a22 | 1.00 | 0.83 | 0.01 | 80.03 |

**Figure 25**     2-color QC Report—Agilent SpikeIns Signal Statistics

**1-color gene expression spike-in signal statistics**

For each sequence of spike-ins this table shows the Probe Name, the median Processed Signal (median of LogProcessedSignal), %CV (SD_ProcessedSignals/Avg_ProcessedSignals) and StdDev (of LogProcessedSignals).

| Agilent SpikeIns Signal Statistics | | | | |
| --- | --- | --- | --- | --- |
| Probe Name | Log (Relative Conc.) | Median (Log Proc. Sig.) | % CV | StdDev |
| (+)E1A_r60_3 | 0.30 | 0.51 | 56.49 | 0.18 |
| (+)E1A_r60_a104 | 1.30 | 0.98 | 17.59 | 0.08 |
| (+)E1A_r60_a107 | 2.30 | 1.67 | 9.72 | 0.04 |
| (+)E1A_r60_a135 | 3.30 | 2.70 | 3.93 | 0.02 |
| (+)E1A_r60_a20 | 3.83 | 3.06 | 5.47 | 0.02 |
| (+)E1A_r60_a22 | 4.30 | 3.63 | 6.56 | 0.03 |
| (+)E1A_r60_a97 | 4.82 | 4.17 | 6.17 | 0.03 |
| (+)E1A_r60_n11 | 5.30 | 4.67 | 6.13 | 0.03 |
| (+)E1A_r60_n9 | 5.82 | 4.69 | 9.38 | 0.04 |
| (+)E1A_r60_1 | 6.30 | 4.81 | 5.69 | 0.02 |

**Figure 26**     1-color QC Report—Agilent SpikeIns Signal Statistics

# Spike-in Linearity Check for 2-color Gene Expression

Using the data calculated for the above table, the observed average log ratio is plotted vs. the expected log ratio for each of the spike-in probes. A linear regression analysis is done using these values and the metrics are shown below the plot. A slope of 1, y-intercept of 0 and $R^2$ of 1 is the ideal of such a linear regression. A slope < 1 may indicate compression, such as having under-corrected for background. The regression coefficient ($R^2$) reflects reproducibility.

The standard deviation for each data point is shown on the plot by an error bar extending above and below the point.
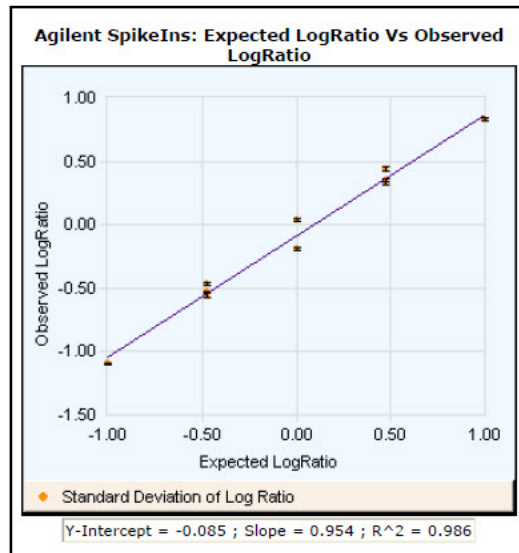
**Figure 27** QC Report—Agilent SpikeIns: Expected Log Ratio Vs. Observed LogRatio

# Spike-in Linearity Check for 1-color Gene Expression

*This plot is usually sigmoidal with two asymptotes, one at the scanner saturation point and one at the level of signal for sequences with no specifically bound target. Some microarrays produce plots missing the top asymptote, especially if extended dynamic range is used. (See the plot below.)*

This plot shows the dose/response curve of the spike-ins from the detection limit to the saturation point.

At high signal levels the error bars are small since the scanner reaches saturation at this point. Both the signals and standard deviations are underestimated because the saturated data is not excluded from the calculation.

At low signal levels the error bars are visible because the signal is dropping into the background noise. The signal level at the top of the error bars of the features with lowest signal provides a rough estimate of the lower limit of detection. Signals at this level can be slightly overestimated and the error slightly underestimated because the signals below zero are excluded from the calculation.

The most reliable Feature Extraction data is found in the signal range where the signal increases linearly with the concentration of the target.
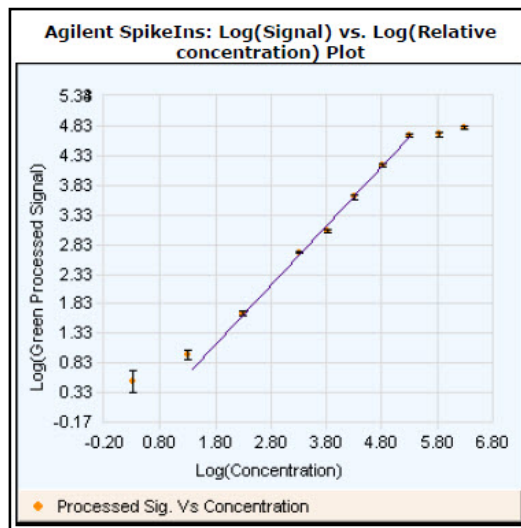


**Figure 28**   1-color QC Report—Agilent SpikeIns: Log (Signal) vs. Log (Relative concentration) Plot

## Table of Values for Concentration-Response Plot (1-color only)

This table presents the values for the log signal vs. log concentration plot shown in Figure 28.



**Agilent Spike-In Concentration-Response Statistics**
**Linear Range Statistics:**

| | |
|---|---|
| Low Signal | 0.73 |
| High Signal | 4.67 |
| Low Relative Concentration | 1.38 |
| High Relative Concentration | 5.33 |
| Slope | 1.00 |
| R^2 Value | 1.00 |

**Signal Detection Limit Statistics**

| | |
|---|---|
| Saturation Point | 4.81 |
| Low Threshold | 0.51 |
| Low Threshold Error | 0.20 |
| Spike-In Detection Limit | 1.05 |

**Figure 29**    1-color QC Report—Agilent Spike-In Concentration-Response Statistics

## Detection of missing spike-ins

This section describes how Feature Extraction deals with missing spike-ins.

Case 1. If the array has a Grid Template with NO SpikeIns in the design,

- If standard protocol is run, then Feature Extraction will give a Warning in the Summary Report that there are no SpikeIn probes.
- If protocol has "SpikeIn Used" set to False, then the QC metric table in the QC Report will show "-" for values, and black font (instead of red, green, or blue fonts) indicating no evaluation has been done by Feature Extraction. Specialized SpikeIn plots & tables will be omitted from the report.

Case 2. If the array has a Grid Template WITH SpikeIns in the design, but the user adds no SpikeIns to hyb,

- If standard protocol is run, the results will either be wrong values or listed as "NA".
- If the protocol has "SpikeIn Used" set to False then the QC metric table in the QC Report will show "-" for values, and black font (instead of red, green, or blue fonts) indicating no evaluation has been done by Feature Extraction. Specialized SpikeIn plots & tables will be omitted from the report.

### How the curve and statistics are calculated

**Curve fit equation**    All of the statistics in the table above are calculated using a parameterized sigmoidal curve fit to the data.

$$F(x) = min + \frac{max - min}{1 + e^{(-(x - x0))/w}}$$

where $min$ is the level of signal for sequences with no specifically bound target and $max$ is the upper limit of detection

where $x0$ is the center of the data and close to the center of the linear range

where $w$ is the width of the curve on either side of $x0$.

**Curve fit calculations**    Before the calculations the following assumptions are made:

- Saturation Point is fixed or close to scanner detection limit. This value is Log(Scanner Saturation Value) = 4.82.
- The linear range of the curve, (x0-w) (x0+w), does not define the dynamic range of the data as the data is close to linear for higher multiples of w away from x0.

- The asymptotes for the max and the min are not necessarily symmetric. The upper asymptote is a function of scanner offset, and the lower asymptote is a function of chemistry/scanner noise.

The calculations then follow this order:

**a** The Min is estimated by taking all the SpikeIn data and for each sequence calculating the BackgroundSubtracted-SignalAverage, the Median of the Log of the processed Signals, StDev of the Log of the processed Signals, the %CV of the processed signals.

The Median Log Proc Signal, %CV, StDev of the Log of the processed signals all show up in the Agilent SpikeIns Signal Statistics table of the QC report.

For each sequence, use the calculated Background-SubtractedSignalAverage and compare against the StdDeviation of the Negative Controls (StdDevBgSubSigNegCtrl) using the formula BGSubAverage * MultErrorGreen > StdDevBgSubSigNegCtrl. Exclude the Proc Signals that fail this test, and use the median of the Proc Signals for the remaining sequences as the initial guess.

**b** Max is estimated as Log(Scanner SaturationValue).

**c** x0 is estimated by starting with the y-value (max+min)/2, then finding the 2 closest Med Log Proc Signals above and below this point. Finding the Log(concentrations) of those points and then computing a slope and an intercept by

slope = (MedianLogProcSig[HIGH] – MedianLogProcSig[LOW])/(LogConc[HIGH] – LogConc[LOW]); intercept = LogConc[HIGH] – slope * MedianLogProcSig[HIGH]

**d** w is estimated by using the slope calculated above. By looking at the derivative of F(x) at x0 we get DF(x):x0 = (max-min)/4*w so w = 4*slope / (max – min).

**e** After the estimates are complete the data is fit and the parameters (Min,Max, x0, w) are optimized by using a parameterized curve fitting routine (called

Levenberg-Marquardt and is a standard technique documented in Numerical Recipes in C on pages 683 – 688).

**f** After the curve fitting is done, the Low Relative Concentration is calculated as x0 – 2.3*w.

**g** The High relative Concentration is calculated as x0 + 2.2*w.

**h** All the eQC points falling between x0 – 2.3*w and x0 + 2.2*w are then fit through a line with the Slope and R-Squared value reported.

**i** All of the points with a concentration below Low Concentration are used to calculate SpikeIn Detection limit. For each probe, the mean and standard deviation is calculated in linear BGSubSignal space. Then the average plus 1 standard deviation is calculated for each probe. The maximum of these is used. It is converted to log10 space and reported as the SpikeIn Detection Limit.

**Relation of curve fit calculations to statistics in table**  In summary, the table below presents descriptions of the statistics in Figure 29, their definitions within the equation and their output in the stats table.

**Table 2**   Spike-In Concentration-Response Statistics for 1-color microarrays

| Statistic | Description | Where in calculations | Stats Table Output |
|---|---|---|---|
| Saturation Point | upper limit of detection | max-step b | eQCOneColorLogHighSignal |
| Low Threshold | lower limit of detection | min-step a | eQCOneColorLogLowSignal |
| Low Threshold Error | error for lower limit | See equation below table | eQCOneColorLogLowSignalError |
| Low Signal | lowest quantifiable signal in linear range | lowest signal from linear fit in step h | eQCOneColorLinFitLogLowSignal |
| High Signal | highest quantifiable signal in linear range | highest signal from linear fit in step h | eQCOneColorLinFitLogHighSignal |
| Low Relative Concentration | lowest concentration leading to quantifiable signal | x0-2.3w in step f | eQCOneColorLinFitLogLowConc |

**Table 2**    Spike-In Concentration-Response Statistics for 1-color microarrays

| Statistic | Description | Where in calculations | Stats Table Output |
|---|---|---|---|
| High Relative Concentration | highest concentration leading to quantifiable signal | x0+2.2w in step g | eQCOneColorLinFitLogHighConc |
| Slope | slope of the linear fit on sigmoidal curve | from step h | eQCOneColorLinFitSlope |
| R^2 Value | correlation coefficient for linear fit | from step h | eQCOneColorLinFitRSQ |
| SpikeIn Detection Limit | The average plus 1 standard deviation of the spike ins below the linear concentration range | from step i | eQCOneColorSpikeInDetectionLimit |

$$LowThresholdError = \sqrt{\sum_A SD(Log(ProcessedSignals))^2}$$

where the set $A$ is from step a in the table

**Accuracy of linear fit to middle of sigmoidal curve**   Agilent calculated the % difference between expected log processed signals at the high and low relative concentrations on the linear curve with the expected log signals for the same concentrations on the sigmoidal curve.

For the high end of the linear range, the % difference is 15.36%.

For the low end of the linear range, the % difference is 16.75%.

# QC Report Results in the FEPARAMS and Stats Tables

The FEPARAMS table contains most of the QC header information. The Stats table output contains all the metrics shown on the QC Reports. These QC stats let you make "tracking" charts of individual metrics that you may want to follow over time. To separate out the FEPARAMS and Stats tables from each other and the FEATURES table, see the *Agilent Feature Extraction for CytoGenomics User Guide*.

# QC Metric Set Results

The figures below show the metric names and default thresholds for the QC metric set results that appear in the Evaluation Tables for each of the QC metric sets available for Feature Extraction for CytoGenomics.

*You can display the QC Metric Set Properties by double-clicking on a QC metric set in the QC Metric Set Browser.*

For details on the logic used for evaluating metrics, see "Metric Evaluation Logic" on page 66.

Note that SNP probes are not used in calculation of any CGH QC Metric.

# CytoCGH_QCMT_1x_Nov17

QC Metric Set Properties - CytoCGH_QCMT_1x_Nov17

**Metric Set Info**

| | |
|---|---|
| Created On: | 13-Nov-2017 11:01 |
| Type: | Agilent |
| Description: | |
| Removable : | False |

**Metrics**

| Metric Name | Excellent | Good | Evaluate |
|---|---|---|---|
| IsGoodGrid | >1 | NA | <1 |
| AnyColorPrcntFeatNonU... | <1 | 1 to 5 | >5 |
| DerivativeLR_Spread | <0.20 | 0.20 to 0.30 | >0.30 |
| gRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| g_BGNoise | <8 | 8 to 15 | >15 |
| g_Signal2Noise | >100 | 30 to 100 | <30 |
| g_SignalIntensity | >150 | 50 to 150 | <50 |
| rRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| r_BGNoise | <8 | 8 to 15 | >15 |
| r_Signal2Noise | >100 | 30 to 100 | <30 |
| r_SignalIntensity | >150 | 50 to 150 | <50 |
| RestrictionControl | | 0.80 to 1 | <0.80 or >1 |
| LogRatioImbalance | -0.26 to 0.26 | (-0.75 to -0.26) or (0.26 to 0.7... | <-0.75 or >0.75 |

Close

**Figure 30**    QC Metrics for CytoCGH_QCMT_1x_Nov17 metric set

# CytoCGH_QCMT_2x_Nov17

QC Metric Set Properties - CytoCGH_QCMT_2x_Nov17

**Metric Set Info**

| | |
|---|---|
| Created On: | 13-Nov-2017 11:01 |
| Type: | Agilent |
| Description: | |
| Removable : | False |

**Metrics**

| Metric Name | Excellent | Good | Evaluate |
|---|---|---|---|
| IsGoodGrid | >1 | NA | <1 |
| AnyColorPrcntFeatNonU... | <1 | 1 to 5 | >5 |
| DerivativeLR_Spread | <0.20 | 0.20 to 0.30 | >0.30 |
| gRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| g_BGNoise | <10 | 10 to 20 | >20 |
| g_Signal2Noise | >100 | 50 to 100 | <50 |
| g_SignalIntensity | >300 | 100 to 300 | <100 |
| rRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| r_BGNoise | <10 | 10 to 20 | >20 |
| r_Signal2Noise | >100 | 50 to 100 | <50 |
| r_SignalIntensity | >300 | 100 to 300 | <100 |
| RestrictionControl | | 0.80 to 1 | <0.80 or >1 |
| LogRatioImbalance | -0.26 to 0.26 | (-0.75 to -0.26) or (0.26 to 0.7... | <-0.75 or >0.75 |

Close

**Figure 31**    QC Metrics for CytoCGH_QCMT_2x_Nov17 metric set

# CytoCGH_QCMT_4x_Nov17



QC Metric Set Properties - CytoCGH_QCMT_4x_Nov17

**Metric Set Info**

| | |
|---|---|
| Created On: | 13-Nov-2017 11:01 |
| Type: | Agilent |
| Description: | |
| Removable : | False |

**Metrics**

| Metric Name | Excellent | Good | Evaluate |
|---|---|---|---|
| IsGoodGrid | >1 | NA | <1 |
| AnyColorPrcntFeatNonU... | <1 | 1 to 5 | >5 |
| DerivativeLR_Spread | <0.20 | 0.20 to 0.30 | >0.30 |
| gRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| g_BGNoise | <10 | 10 to 20 | >20 |
| g_Signal2Noise | >100 | 50 to 100 | <50 |
| g_SignalIntensity | >400 | 200 to 400 | <200 |
| rRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| r_BGNoise | <10 | 10 to 20 | >20 |
| r_Signal2Noise | >100 | 50 to 100 | <50 |
| r_SignalIntensity | >400 | 200 to 400 | <200 |
| RestrictionControl | | 0.80 to 1 | <0.80 or >1 |
| LogRatioImbalance | -0.26 to 0.26 | (-0.75 to -0.26) or (0.26 to 0.7... | <-0.75 or >0.75 |

Close

**Figure 32** QC Metrics for CytoCGH_QCMT_4x_Nov17 metric set

# CytoCGH_QCMT_8x_Nov17



QC Metric Set Properties - CytoCGH_QCMT_8x_Nov17

**Metric Set Info**

Created On: 13-Nov-2017 11:01
Type: Agilent
Description:
Removable : False

**Metrics**

| Metric Name | Excellent | Good | Evaluate |
|---|---|---|---|
| IsGoodGrid | >1 | NA | <1 |
| AnyColorPrcntFeatNonU... | <1 | 1 to 5 | >5 |
| DerivativeLR_Spread | <0.20 | 0.20 to 0.30 | >0.30 |
| gRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| g_BGNoise | <15 | 15 to 25 | >25 |
| g_Signal2Noise | >100 | 30 to 100 | <30 |
| g_SignalIntensity | >400 | 200 to 400 | <200 |
| rRepro | 0 to 0.10 | 0.10 to 0.20 | <0 or >0.20 |
| r_BGNoise | <15 | 15 to 25 | >25 |
| r_Signal2Noise | >100 | 30 to 100 | <30 |
| r_SignalIntensity | >400 | 200 to 400 | <200 |
| RestrictionControl | | 0.80 to 1 | <0.80 or >1 |
| LogRatioImbalance | -0.26 to 0.26 | (-0.75 to -0.26) or (0.26 to 0.7... | <-0.75 or >0.75 |

Close

**Figure 33**    QC Metrics for CytoCGH_QCMT_8x_Nov17 metric set

# CytoCGH_QCMT_SingleCell_Nov17



**Figure 34**     QC Metrics for CytoCGH_QCMT_SingleCell_Nov17 metric set

# Metric Evaluation Logic

When a QC metric set is associated with a protocol, it is used to evaluate results using up to three defined threshold values for given metrics. Results are then flagged in the QC Report Evaluation Metrics table according to the logic described in the following diagram and tables.

Figure 35 shows the metric evaluation using three threshold levels. The black dots indicate how a result is evaluated if its value is the same as a limit value.
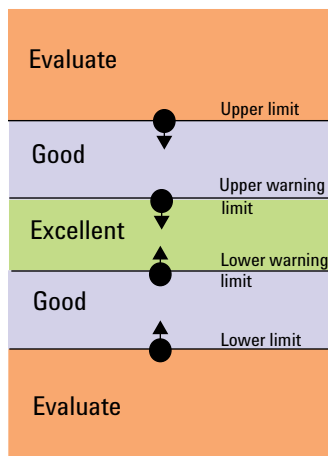


**Figure 35**    Three-level QC Metrics evaluation used for Feature Extraction

The following tables describe how results are evaluated using up to three threshold levels.

### Metric Evaluation Logic tables

In the following tables, evaluation metrics are described for 18 cases (IDs). Results are compared to four limit values, shown in the "Limits used" table: upper limit, upper warning limit, lower warning limit, and lower limit (v1 through v4). The logic used is described in the center table, showing the metric evaluation indication (Excellent, Good, Evaluate) that is based on how the result compares to the given limit

value(s). Cases covered indicate the type of threshold along with the boundaries that are displayed in the QC Report.

(value > Upper limit)  => Evaluate
(value > Upper Warning limit) and (value <= Upper limit)  => Good
(value >= Lower Warning limit) and (value <= Upper warning limit) => Excellent
(value >= Lower limit) and (value < Lower Warning limit) => Good
(value < Lower limit) => Evaluate

**Limits used:**

| ID | Upper limit | Upper Warning limit | Lower Warning limit | Lower limit |
|----|----|----|----|----|
| 1 | v1 | | | |
| 2 | v1 | | | v4 |
| 3 | | | | v4 |
| 4 | | v2 | | |
| 5 | | v2 | v3 | |
| 6 | | | v3 | |
| 7 | v1 | v2 | | |
| 8 | v1 | v2 | v3 | v4 |
| 9 | | | v3 | v4 |
| 10 | v1 | v2 | | |
| 11 | v1 | v2 | v3 | v4 |
| 12 | | | v3 | v4 |
| 13 | v1 | v2 | v3 | |
| 14 | v1 | v2 | | v4 |
| 15 | v1 | | v3 | |
| 16 | v1 | | v3 | v4 |
| 17 | | v2 | | v4 |
| 18 | | v2 | v3 | v4 |

**Logic used:**

| Excellent | Good | Evaluate |
|----|----|----|
| | <= v1 | > v1 |
| | v1 to v4 | > v1 or < v4 |
| | >=v4 | < v4 |
| <= v2 | > v2 | |
| v2 to v3 | > v2 or <v3 | |
| >= v3 | < v3 | |
| <= v2 | NA (special case when v1 = v2) | > v1 |
| v2 to v3 | NA (special case when v1 = v2; and v3 = v4) | > v1 or < v4 |
| >= v3 | NA (special case when v3 = v4) | <v4 |
| <= v2 | v1 to v2 | > v1 |
| v2 to v3 | (v1 to v2) or (v3 to v4) | > v1 or < v4 |
| >= v3 | v3 to v4 | < v4 |
| v2 to v3 | (v1 to v2) or < v3 | > v1 |
| v2 to v4 | (v1 to v2) | > v1 or < v4 |
| v1 to v3 | <v3 | > v1 |
| v1 to v3 | (v3 to v4) | > v1 or < v4 |
| v2 to v4 | > v2 | < v4 |
| v2 to v3 | > v2 or (v3 to v4) | <v4 |

**Cases covered:**

| Type | Boundaries | ID |
|----|----|----|
| **2-level metrics used in FEv10.5:** | | |
| Upper | Good/Evaluate | 1 |
| Range | Good/Evaluate | 2 |
| Lower | Good/Evaluate | 3 |
| **2-level metrics that may be used in FEv10.7 :** | | |
| Upper | Excellent/Good | 4 |
| Range | Excellent/Good | 5 |
| Lower | Excellent/Good | 6 |
| Upper | Excellent/Evaluate | 7 |
| Range | Excellent/Evaluate | 8 |
| Lower | Excellent/Evaluate | 9 |
| **3-level metrics that may be used in FEv10.7 :** | | |
| Upper | Excellent/Good/Evaluate | 10 |
| Range | Excellent/Good/Evaluate | 11 |
| Lower | Excellent/Good/Evaluate | 12 |
| **3-level metrics that are asymmetric; supported but not normally used (FEv10.7):** | | |
| Range | Excellent/Good/Evaluate | 13 |
| Range | Excellent/Good/Evaluate | 14 |
| Range | Excellent/Good/Evaluate | 15 |
| Range | Excellent/Good/Evaluate | 16 |
| Range | Excellent/Good/Evaluate | 17 |
| Range | Excellent/Good/Evaluate | 18 |

**Figure 36**    QC Metrics evaluation tables and cases

# 3
# Text File Parameters and Results

Feature Extraction produces a tab-delimited text file that contains three tables of input parameters and output results.

These tables are FEPARAMS, STATS, and FEATURES. These three tables list all the possible parameters, statistics and feature results that can be generated in the text output file.

**FEPARAMS table**   Contains input parameters and options used to run Feature Extraction.

**STATS table**   Gives results derived from statistical calculations that apply to all features on the microarray.

**FEATURES table**   Displays results for each feature in over 90 output columns, such as gene name, log ratio, processed signal, mean signal, or dye-normalized signal.

**Agilent Technologies**

You have the option in the Project Properties sheet of selecting to generate either the FULL set of parameters, statistics and feature information, COMPACT, QC or MINIMAL. COMPACT output package is the default.

The COMPACT output package contains only those columns that are required by GeneSpring and DNA Analytics software. The tables on the following pages present the text file summary for all output package types (FULL, COMPACT, QC, or MINIMAL).

NOTE   Some of the parameters, statistical results, and feature results may not be included from any one output file, depending on the application and protocol used for Feature Extraction.

You also have the option to generate one file with all three tables or three separate files with one for each table. To select to generate one file or three, see the *Agilent Feature Extraction for CytoGenomics User Guide*.

To display the text results file in an easy-to-read format, see the *Agilent Feature Extraction for CytoGenomics User Guide*.

# Parameters/options (FEPARAMS)

The top-most section of the result file contains the parameters and option choices that you used to run Feature Extraction.

## FULL FEPARAMS Table

**Table 3**    List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Protocol _Name | text | Name of protocol used |
| | Protocol_date | text | Date the protocol was last modified |
| | Scan_date | text | Date the image was scanned |
| | Scan_ScannerName | text | Serial number of the scanner used |
| | Scan_NumChannels | integer | Number of channels in the scan image |
| | Scan_MicronsPerPixelX | float | Number of microns per pixel in the X axis of the scan image |
| | Scan_MicronsPerPixelY | float | Number of microns per pixel in the Y axis of the scan image |
| | Scan_OriginalGUID | text | The global unique identifier for the scan image |
| | Grid_Name | text | Grid template name or grid file name |
| | Grid_Date | integer | Date the grid template or grid file was created |
| | Grid_NumSubGridRows | integer | Number of subgrid columns |
| | Grid_NumSubGridCols | integer | Number of subgrid columns |
| | Grid_NumRows | integer | Number of spots per row of each subgrid |
| | Grid_NumCols | integer | Number of spots per column of each subgrid |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Grid_RowSpacing | float | Space between rows on the grid |
| | Grid_ColSpacing | float | Space between column on the grid |
| | Grid_OffsetX | float | In a dense pack array, the offset in the X direction |
| | Grid_OffsetY | float | In a dense pack array, the offset in the Y direction |
| | Grid_NomSpotWidth | float | Nominal width in microns of a spot from grid |
| | Grid_NomSpotHeight | float | Nominal height in microns of a spot from grid |
| | Grid_GenomicBuild | text | The build of the genome used to create the annotation (if available). If the genome build is not available (not all designs have this information), then it is not put out. All recent and all future designs have it. |
| | FeatureExtractor_Barcode | text | Barcode of the Agilent microarray read from the scan image |
| | FeatureExtractor_Sample | text | Names of hybridized samples (red/green) |
| | FeatureExtractor_ScanFileName | text | Name of the scan file used for Feature Extraction |
| | FeatureExtractor_ArrayName | text | Microarray filename |
| | FeatureExtractor_DesignFileName | text | Design or grid file used for Feature Extraction |
| | FeatureExtractor_PrintingFileName | text | Print file (if available) used for Feature Extraction |
| | FeatureExtractor_PatternName | text | Agilent pattern file name |
| | FeatureExtractor_ExtractionTime | text | Time stamp at the beginning of Feature Extraction run for the extraction set |
| | FeatureExtractor_UserName | text | Windows Log-In Name of the User who ran Feature Extraction |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | FeatureExtractor_ComputerName | text | Computer name on which Feature Extraction was run |
| | FeatureExtractor_ScanFileGUID | text | GUID of the scan file |
| | FeatureExtractor_IsXDRExtraction | integer<br>1 = True<br>0 = False | Indicates whether or not the extraction was an XDR extraction. |
| | DyeNorm_NormFilename | text | Name of the dye normalization list file |
| | DyeNorm_NormNumProbes | integer | Number of probes in the dye normalization list |
| | Grid_IsGridFile | boolean | Indicates whether the grid is from a grid file. |
| | Scan_NumScanPass | 1 or 2 | For 5 micron scans, indicates whether the scan mode was a single (1) or double-pass scan mode on the Agilent Scanner. |
| Place Grid | GridPlacement_Version | text | Version of the grid placement algorithm |
| Place Grid | GridPlacement_ArrayFormat | integer | Choices for grid placement based on the format of the image. Choices include:<br>Automatically Determine<br>Single Density (11k, 22k)<br>Double Density (44k)<br>95k<br>185 (5 and 10 uM)<br>65 micron (5 and 10 uM)<br>30 micron single pack<br>30 micron multi pack<br>244 (5 and 10 uM)<br>25k<br>Third Party |
| Place Grid | GridPlacement_enableOriginXCal | integer<br>1 = True<br>0 = False | Indicates status of the Use the correlation method to obtain origin X of subgrids flag |

**Table 3**   List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Place Grid | GridPlacement_enableUseCentralPack | integer<br>1 = True<br>0 = False | Indicates status of the Use central part of pack for slope and skew calculation flag |
| Place Grid | GridPlacement_placementMode | integer<br><br>0<br>1 | Mode of grid placement<br><br>Allow the grid to distort<br>Place the grid rigidly allowing only translation and rotation |
| Optimize Grid Fit | IterativeSpotFind_CornerAdjust | integer<br>0 = False<br>1 = True | Indicates whether or not the grid will be adjusted for better fit by looking at corner spots on the microarray |
| Optimize Grid Fit | IterativeSpotFind_AdjustThreshold | float | Grid will be adjusted if absolute average difference between grid and spot positions is greater than this fraction |
| Optimize Grid Fit | IterativeSpotFind_MaxIterations | integer | Maximum number of times spot finder algorithm is run to optimize the grid fit |
| Optimize Grid Fit | IterativeSpotFind_FoundSpot Threshold | float | Grid will be adjusted if this fraction or more of the features are considered found by the spot finder algorithm |
| Optimize Grid Fit | IterativeSpotFind_NumCornerFeatures | integer | Indicates the square area of features in each corner of the microarray to be used to calculate the average difference |
| Find Spots | SpotAnalysis_Version | text | Version of the spot analysis algorithm |
| Find Spots | SpotAnalysis_weakthresh | float | Minimum difference between the average intensities of feature and background after Kmeans Initialization |
| Find Spots | SpotAnalysis_MinimumNumPixels | integer | Minimum number of pixels required for the spot analysis |
| Find Spots | SpotAnalysis_RegionOfInterest Multiplier | float | Multiplier that defines how big the Region of Interest (ROI) is in terms of nominal spot spacing |
| Find Spots | SpotAnalysis_convergence_factor | float | Convergence factor of KMeans algorithm |

**Table 3**    List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Find Spots | SpotAnalysis_max_em_iter | integer | Maximum number of iterations of the Bayesian Classification |
| Find Spots | SpotAnalysis_max_reject_ratio | float | Maximum fraction of pixels to be rejected while software performs spotfinding |
| Find Spots | SpotAnalysis_kmeans_rad_reject_factor | float | Factor that defines how much individual spot size may vary relative to the nominal spot size |
| Find Spots | SpotAnalysis_kmeans_cen_reject_factor | float | Factor that defines how far the actual centroid may move relative to its nominal grid position (in terms of nominal radius). In the protocol this parameter is called the Spot Deviation Limit. |
| Find Spots | SpotAnalysis_kmeans_moi_reject_factor | float | Maximum allowable moment of inertia of the spot |
| Find Spots | SpotAnalysis_isspot_factor | float | Factor from the statistics of the found feature and background that indicates if the spot is a spot. |
| Find Spots | SpotAnalysis_isweakspot_factor | float | Factor from the statistics of the found feature and background that indicates if the spot is a strong one. |
| Find Spots | SpotAnalysis_BackgroundThreshold | float | Factor by which the individual spot background may vary from the running average of all the background means. |
| Find Spots | SpotAnalysis_ROIType | integer | Type of Region of Interest |
| Find Spots | SpotAnalysis_UseNominalDiameterFromGT | integer 1 = True 0 = False | If True, the nominal spot diameter from the grid template is used as a starting point for final spot diameter computation. If False, the nominal diameter is obtained from the grid placement algorithm. |
| Find Spots | SpotAnalysis_RejectMethod | integer 0 2 3 | Pixel Outlier Rejection turned off Standard Deviation based Interquartile Range based |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Find Spots | SpotAnalysis_StatBoundFeat | float | Multiplier parameters for feature outlier rejection method as selected above |
| Find Spots | SpotAnalysis_StatBoundBG | float | Multiplier parameters for background outlier rejection method as selected above |
| Find Spots | SpotAnalysis_SpotStatsMethod | integer<br>1<br>2 | Different algorithms to calculate spot statistics<br>CookieCutter method<br>Whole Spot method |
| Find Spots | SpotAnalysis_CookiePercentage | float | The fraction of the nominal radius used to draw the cookie around the centroid of each spot |
| Find Spots | SpotAnalysis_ExclusionZone Percentage | float | The outer radius of the exclusion zone based on nominal spot size |
| Find Spots | SpotAnalysis_EstimateLocalRadius | integer<br>1 = True<br>0 = False | The option to calculate the outer radius of the local background based on row and column spacing |
| Find Spots | SpotAnalysis_LocalBGRadius | float | The outer radius of the local background supplied from the protocol if EstimateLocalRadius is not selected |
| Find Spots | SpotAnalysis_SignalMethod | integer | The option for the statistical method for determining signals from features: either mean (and standard deviation) or median (and normalized IQR).<br><br>Mean is 1 and Median is 2. |
| Find Spots | SpotAnalysis_ComputePixelSkew | integer<br>true = 1<br>false = 0 | The option to set whether the program computes and shows the skew of each feature. Default is false. |
| Find Spots | SpotAnalysis_PixelSkewCookiePct | float<br>(0.00-1.00;<br>0.70 default) | The percentage of the feature that should be used when calculating the pixel skew. A value of .70 means 70% of the radius of the feature. |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Find Spots | SpotAnalysis_CentroidDiff | Integer<br>1 = True<br>0 = False | The software computes the per feature Centroid Difference between the Grid position and the Spot Center. |
| Find Spots | SpotAnalysis_NozzleAdjust | Integer<br>1 = True<br>0 = False | The software attempts to adjust a nozzle group in order to compensate for variations in printing. |
| Flag Outliers | OutlierFlagger_Version | text | Version of Outlier Flagger algorithm |
| Flag Outliers | OutlierFlagger_NonUnifOLOn | integer<br>1 = True<br>0 = False | NonUniformity Outlier flagging turned on<br>NonUniformity Outlier flagging turned off |
| Flag Outliers | OutlierFlagger_FeatATerm | float | Applies to feature: specifies the intensity dependent variance and is set to the square of the CV |
| Flag Outliers | OutlierFlagger_FeatBTerm | float | Applies to feature: specifies the variance due to the Poisson distributed noise |
| Flag Outliers | OutlierFlagger_FeatCTerm | float | Applies to feature: specifies variance due to background noise of the scanner, slide glass, and other signal-independent sources |
| Flag Outliers | OutlierFlagger_BGATerm | float | Applies to background: specifies the intensity-dependent variance and is set to the square of the CV |
| Flag Outliers | OutlierFlagger_BGBTerm | float | Applies to background: specifies the variance due to the Poisson distributed noise |
| Flag Outliers | OutlierFlagger_BGCTerm | float | Applies to background: specifies variance due to background noise of the scanner, slide glass, and other signal-independent sources |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Flag Outliers | OutlierFlagger_OLAutoComputeABC | integer<br>1 = True<br>0 = False | AutoCompute Outlier flagging turned on<br>AutoCompute Outlier flagging turned off<br><br>For Agilent protocols when this flag is turned on, the polynomial is calculated automatically. This means that all above Feature and BG terms for B and C no longer appear in the output. Rather, they are calculated automatically and appear in the STATS table. Also, the eight parameters following this row appear. |
| Flag Outliers | OutlierFlagger_FeatBCoeff | float | Feature: Red Poissonian Noise Term Multiplier |
| Flag Outliers | OutlierFlagger_FeatCCoeff | float | Feature: Red Signal Constant Term Multiplier |
| Flag Outliers | OutlierFlagger_FeatBCoeff2 | float | Feature: Green Poissonian Noise Term Multiplier |
| Flag Outliers | OutlierFlagger_FeatCCoeff2 | float | Feature: Green Signal Constant Term Multiplier |
| Flag Outliers | OutlierFlagger_BGBCoeff | float | Background: Red Poissonian Noise Term Multiplier |
| Flag Outliers | OutlierFlagger_BGCCoeff | float | Background: Red Signal Constant Term Multiplier |
| Flag Outliers | OutlierFlagger_BGBCoeff2 | float | Background: Green Poissonian Noise Term Multiplier |
| Flag Outliers | OutlierFlagger_BGCCoeff2 | float | Background: Green Signal Constant Term Multiplier |
| Flag Outliers | OutlierFlagger_PopnOLOn | integer<br>1 = True<br>0 = False | Population Outlier flagging turned on<br>Population Outlier flagging turned off |
| Flag Outliers | OutlierFlagger_MinPopulation | integer | Minimum number of replicates to turn on population outlier flagging |

**Table 3**    List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Flag Outliers | OutlierFlagger_IQRatio | float | The boundary conditions for conducting box-plot analysis to isolate population outliers |
| Flag Outliers | OutlierFlagger_BackgroundIQRatio | float | The boundary conditions for conducting box-plot analysis to isolate population outliers for the background |
| Flag Outliers | OutlierFlagger_Use Qtest | integer<br>1 = True<br>0 = False | Enables Qtest statistics when the minimum number of replicates for population outliers is greater than 2 and less than the minimum population specified in the outlier section of the protocol. |
| Flag Outliers | OutlierFlagger_UsePopnOLInMAGE | integer<br>1 = True<br>0 = False | Indicates whether to report population outliers as "Failed" in MAGEML output |
| Compute Bkgd, Bias and Error | BGSubtractor_MultiplicativeDetrend On | integer<br>1 = True<br>0 = False | Enables multiplicative detrending. 1-color and CGH microarray protocols have this parameter enabled. |
| Compute Bkgd, Bias and Error | BGSubtractor_MultDetrendWinFilter | integer<br>0<br>1<br>2 | No filtering<br>Average filtering<br>Median filtering |
| Compute Bkgd, Bias and Error | BGSubtractor_MultDetrendIncrement | integer | The increment in number of features by which the square window is shifted horizontally and vertically on the microarray. |
| Compute Bkgd, Bias and Error | BGSubtractor_MultDetrendWindow | integer | Specifies size of the square window by the number of rows and columns. The specified percentage of low intensity features is selected from this window size. |
| Compute Bkgd, Bias and Error | BGSubtractor_MultDetrendNeighbor-hoodSize | float<br>[0-1] | Specifies the fraction of total number of neighborhood data points that will be weighted for linear regression during surface fitting for each data point |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Compute Bkgd, Bias and Error | BGSubtractor_MultHighPassFilter | integer<br>1 = True<br>0 = False | Enables rejection of probes close to zero signal from the set of features used in the fit. |
| Compute Bkgd, Bias and Error | BGSubtractor_PolynomialMultiplicativeDetrend | integer<br>1 = True<br>0 = False | The option to use a polynomial surface fit method for the multiplicative detrending fit (rather than LOESS). |
| Compute Bkgd, Bias and Error | BGSubtractor_NegCtrlThresholdMultDetrendFactor | float | This factor multiplies the negative control spread to determine the threshold signal below which low intensity features are filtered out of the multiplicative detrending fit set. |
| Compute Bkgd, Bias and Error | BGSubtractor_PolynomialMultiplicativeDetrendDegree | integer<br>[-1, 5] | Shows the degree of the polynomial fit used for the multiplicative detrending. The most common choices are 2 (quadratic or 2nd order surface) and 4 (4th order surface). |
| Compute Bkgd, Bias and Error | BGSubtractor_TestMultDetrendOnCVs | integer | Tests whether the replicate CVs improve (i.e. decrease) after multiplicative detrending. If this choice is 1=True, and the replicate CVs don't improve, Feature Extraction doesn't use the multiplicative detrending for that array. |
| Compute Bkgd, Bias and Error | BGSubtractor_MultDetrendOn Replicates | integer<br>1 = True<br>0 = False | Specifies to use only replicated probes (with multiple features) normalized to their replicate average for the multiplicative detrending set. |

**Table 3**    List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Compute Bkgd, Bias and Error | BGSubtractor_BGSubMethod | integer | |
| | | 1 | Either minimum feature or minimum local background across the microarray for background subtraction (global method) |
| | | 2 | Average of local backgrounds for background subtraction (global method) |
| | | 3 | Average of negative controls for background for background subtraction (global method) |
| | | 5 | Local background corresponding to each feature for background subtraction (local method) |
| | | 6 | Minimum feature across the microarray for background subtraction (global method) |
| | | 7 | No background subtraction |
| Compute Bkgd, Bias and Error | BGSubtractor_MaxPVal | float | The pValue at which a feature is determined to be statistically significant above background |
| Compute Bkgd, Bias and Error | BGSubtractor_WellAboveMulti | float | The number of standard deviations above background at which the feature is flagged as well above background |
| Compute Bkgd, Bias and Error | BGSubtractor_BackgroundCorrectionOn | integer 1 = True 0 = False | Globally adjust background turned on Globally adjust background turned off |
| Compute Bkgd, Bias and Error | BGSubtractor_BgCorrectionOffset | | Adjust the signal of all features by an offset constant so that very low signal features end up at this offset. Appears when Globally adjust background is turned on. |
| Compute Bkgd, Bias and Error | BGSubtractor_CalculateSurface MetricsOn | integer 1 = True 0 = False | Surface fit is done and metrics calculated. Surface fit and metrics are not done. |

**Table 3**    List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Compute Bkgd, Bias and Error | BGSubtractor_SpatialDetrendOn | integer<br>1 = True<br>0 = False | <br>Spatial detrend turned on<br>Spatial detrend turned off |
| Compute Bkgd, Bias and Error | BGSubtractor_DetrendLowPassFilter | integer<br>1 = True<br>0 = False | <br>Low pass filter used<br>Low pass filter not used |
| Compute Bkgd, Bias and Error | BGSubtractor_DetrendLowPass Percentage | integer | Specifies percentage of features based on the lowest intensity probes in each window that will be used to fit the surface |
| Compute Bkgd, Bias and Error | BGSubtractor_DetrendLowPass Window | integer | Specifies size of the square window by the number of rows and columns. The specified percentage of low intensity features is selected from this window size. |
| Compute Bkgd, Bias and Error | BGSubtractor_DetrendLowPass Increment | integer | The increment in number of features by which the above window is shifted horizontally and vertically on the microarray |
| Compute Bkgd, Bias and Error | BGSubtractor_NegCtrlSpreadCoeff | float | The number of multiples of the negative control spread that defines the signal range within which features are considered to be within the negative control range for "FeaturesInNegativeControlRange" background detrend option. |
| Compute Bkgd, Bias and Error | BGSubtractor_NegCtrlSpreadRobust On | float | Specifies to remove negative control features that are outliers before calculating the negative control spread for use with FeaturesInNegativeControlRange. |
| Compute Bkgd, Bias and Error | BGSubtractor_AdditiveDetrend FeatureSet | integer<br><br>0<br>1<br>2 | Determines which features are considered for the surface fit set<br>All inlier features<br>Negative control inliers only<br>Features in negative control range |

**Table 3**    List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Compute Bkgd, Bias and Error | BGSubtractor_DetrendNeighborhood Size | float | Specifies the fraction of total number of neighborhood data points that will be weighted for linear regression during surface fitting for each data point |
| Compute Bkgd, Bias and Error | BGSubtractor_ErrModelSignificance | integer 0 = pixel staistics 1 = error model | Decides whether the error model or pixel staistics are used to determine Positive and Significance calls and WellAboveBackground. |
| Compute Bkgd, Bias and Error | BGSubtractor_RobustNCStats | integer 1 = True 0 = False | Specifies if a variation in the population algorithm is turned on. This algorithm repeats the population outlier IQR algorithm on all features classified as negative controls, after the first pass of population algorithm has been run on each sequence. You may want to use this algorithm when you see "hot" features that have not been flagged as population outliers or "hot" sequences where all features of the sequence have higher signals than those in other negative control sequences. |
| Compute Bkgd, Bias and Error | BGSubtractor_RobustNCOutlierFactor | float | To calculate robust IQR statistics, the algorithm uses upper and lower limits that contain a (Multiplier x IQR) term. This parameter is the Multiplier. |
| Compute Bkgd, Bias and Error | BGSubtractor_ErrorModel | integer  2 0 | Choose universal error, or the most conservative Universal Error Model Most Conservative |
| Compute Bkgd, Bias and Error | BGSubtractor_MultErrorGreen | float | Multiplicative error component in Green channel |
| Compute Bkgd, Bias and Error | BGSubtractor_MultErrorRed | float | Multiplicative error component in Red channel |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Compute Bkgd, Bias and Error | BGSubtractor_AutoEstimateAddErrorGreen | integer<br>1 = True<br>0 = False | <br>Auto-estimation turned on<br>Auto-estimation turned off |
| Compute Bkgd, Bias and Error | BGSubtractor_AutoEstimateAddErrorRed | integer<br>1 = True<br>0 = False | <br>Auto-estimation turned on<br>Auto-estimation turned off |
| Compute Bkgd, Bias and Error | BGSubtractor_AddErrorGreen | float | This additive error component in the green channel is entered in the protocol when auto-estimation is turned off. When auto-estimation is turned on, the estimated error value appears in the Stats table as AddErrorEstimateGreen. |
| Compute Bkgd, Bias and Error | BGSubtractor_AddErrorRed | float | This additive error component in the red channel is entered in the protocol when auto-estimation is turned off. When auto-estimation is turned on, the estimated error value appears in the Stats table as AddErrorEstimateRed. |
| Compute Bkgd, Bias and Error | BGSubtractor_MultNcAutoEstimate | float<br>[0-10] | Multiplier for the first term (standard deviation of the inlier negative control) in the additive error equation. |
| Compute Bkgd, Bias and Error | BGSubtractor_MultRMSAutoEstimate | float<br>[0-10] | Multiplier for the second term (gMultSpatialDetrendRMSFit) in the additive error equation. |
| Compute Bkgd, Bias and Error | BGSubtractor_MultResidualsRMSAutoEstimate | float<br>[0-10] | Multiplier for the third term in the additive error equation. |
| Compute Bkgd, Bias and Error | BGSubtractor_AutoEstimateNCOnlyThresh | float | This parameter is for single density 8-pack microarrays where Feature Extraction may not be able to accurately subtract the background using the spatial detrending method. This parameter provides a minimum number of features needed for the software to use the residual or the RMS to estimate the additive error. It comes up only if using low density 8-pack microarrays. |

**Table 3**    List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Compute Bkgd, Bias and Error | BGSubtractor_UseSurrogates | integer | Flag indicating the use of surrogates |
| | | 1 = True | Use of surrogates turned on |
| | | 0 = False | Use of surrogates turned off |
| Compute Bkgd, Bias and Error | BGSubtractor_Version | text | Version of BGSubtractor algorithm |
| Correct Dye Biases | DyeNorm_Version | text | Version of DyeNorm algorithm |
| Correct Dye Biases | DyeNorm_UseDyeNormList | integer | |
| | | 0 | Automatically determine |
| | | 1 | True |
| | | 2 | False |
| Correct Dye Biases | DyeNorm_SelectMethod | integer | Method for selecting features used for measurement of dye bias: |
| | | 4 | Use All Probes |
| | | 5 | Use List of Normalization Genes |
| | | 6 | Use Rank Consistent Probes |
| | | 7 | Use Rank Consistent List of Normalization Genes |
| Correct Dye Biases | DyeNorm_ArePosNegCtrlsOK | integer | |
| | | 1 = True | Use positive and negative controls for dye normalization. |
| | | 0 = False | Do not use these controls. |
| Correct Dye Biases | DyeNorm_SignalCharacteristics | integer | |
| | | 1 | Only positive and significant signals |
| | | 2 | All positive signals |
| | | 3 | All negative and positive signals |
| Correct Dye Biases | DyeNorm_CorrMethod | integer | Methods for computation of dye normalization factor to remove dye bias |
| | | 0 | Linear |
| | | 1 | Linear&LOWESS (locally weighted linear regression preceded by linear scaling in each dye channel) |
| | | 2 | LOWESS (locally weighted linear regression) |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Correct Dye Biases | DyeNorm_LOWESSSmoothFactor | float | Smoothing parameter (Neighborhood size) for LOWESS curve fitting |
| Correct Dye Biases | DyeNorm_LOWESSNumSteps | integer | Number of iterations in LOWESS |
| Correct Dye Biases | DyeNorm_RankTolerance | float | The threshold to pick rank consistent features between 2 channels for measuring dye biases |
| Correct Dye Biases | DyeNorm_VariableRankTolerance | integer<br>1 = True<br>0 = False | Allows the rank tolerance to vary with signal level to allow a fixed percentage of the data to be considered rank consistent. |
| Correct Dye Biases | DyeNorm_MaxRankedSize | integer | The limit on the number of points used for the dye normalization set. If the number is greater than this, a random subset is chosen using this number of points. |
| Correct Dye Biases | DyeNorm_IsBGPopnOLOn | integer<br>1 = True<br><br>0 = False | Software excludes any features from the dye normalization set if the local backgrounds associated with those features have been flagged as population outliers (in either channel).<br>The default recommendation is False. |
| Compute Ratios | Ratio_Version | text | Version of Ratio algorithm |
| Compute Ratios | Ratio_PegLogRatioValue | float | Both positive and negative log ratio values are capped to this absolute value |
| Calculate Metrics | QCMetrics_UseSpikeIns | integer<br>1 = True<br>0 = False | Use SpikeIns<br>Do not use SpikeIns |
| Calculate Metrics | QCMetrics_minReplicatePopulation | integer | Minimum number of replicates necessary to calculate replicate statistics |
| Calculate Metrics | QCMetrics_differentialExpression PValue | float | The pValue to use to look for differentially expressed genes |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| Calculate Metrics | QCMetrics_MaxEdgeDefect Threshold | float | Maximum allowable fraction of features along any edge of the microarray that are non-uniform before a grid placement warning is given. |
| Calculate Metrics | QCMetrics_MaxEdgeNotFound Threshold | float | Maximum allowable fraction of features along any edge of the microarray that are not found before a grid placement warning is given. |
| Calculate Metrics | QCMetrics_MaxLocalBGNonUnif Threshold | float | Maximum allowable fraction of the local background regions on the microarray that are flagged as NonUniform before a grid placement warning is given. |
| Calculate Metrics | QCMetrics_MinNegCtrlSDev | float | Minimum value for the standard deviation for the negative controls |
| Calculate Metrics | QCMetrics_MinReproducibility | float | Minimum value for the reproducibility |
| Calculate Metrics | QCMetrics_Formulation | integer 1 = TwoColor 2 = OneColor 3 = CGH | The SpikeIn formulation to use for the SpikeIn Calculation. Different formulations will yield different expected values and different concentration values. |
| Calculate Metrics | QCMetrics_EnableDyeFlip | integer 1 = True 2 = False | If True (default), the sign of the slope for the spikeIns plot and its trend will be changed when the slope is detected to have the wrong sign. This means the labelling was intentionally flipped and must be flipped back. |
| Calculate Metrics | QCMetrics_PercentileValuefor Signal | float | The PercentileIntensitySignal is calculated by the software on the [r,g]ProcessedSignal showing the signal at a given percentile over the NonControl features. This parameter is the percentile used for the calculation. By default the value is set to 75; the software generates the 75% Signal value of the ProcessedSignals for all channels available. |
| | FeatureExtractor_Version | text | Version of Feature Extractor |

**Table 3** List of parameters and options contained within the FULL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | FeatureExtractor_SingleTextFile Output | integer<br>1 = True<br><br>0 = False | The system prints the three tables (FEParams, Stats and Features) are printed in the same text file.<br>The system prints each of the three tables in separate text files. |
| | FeatureExtractor_JPEGDownSample Factor | float | Factor by which the image is scaled down and then converted to the JPEG format. Must be at least 2; 1 is no longer allowed. |
| | FeatureExtractor_ColorMode | integer<br>0<br>1<br>2 | A flag to indicate output color<br>One color; green only<br>2-color<br>One color: red only |
| | FeatureExtractor_QCReportType | integer<br>0<br>1<br>2<br>4 | Type of QC report to generate<br>Gene Expression<br>CGH_ChIP<br>miRNA<br>Streamlined CGH |
| | FeatureExtractor_OutputQCReport GraphText | integer<br>1 = True<br>0 = False | Generate output details on QC report graphs |

# COMPACT FEPARAMS Table

**Table 4**    List of parameters and options contained within the COMPACT text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Protocol _Name | text | Name of protocol used |
| | Protocol_date | text | Date the protocol was last modified |
| | Scan_ScannerName | text | Agilent scanner serial number used |
| | Scan_NumChannels | integer | Number of channels in the scan image |
| | Scan_date | text | Date the image was scanned |
| | Scan_MicronsPerPixelX | float | Number of microns per pixel in the X axis of the scan image |
| | Scan_MicronsPerPixelY | float | Number of microns per pixel in the Y axis of the scan image |
| | Scan_OriginalGUID | text | The global unique identifier for the scan image |
| | Scan_NumScanPass | 1 or 2 | For 5 micron scans, indicates whether the scan mode was a single (1) or double-pass scan mode on the Agilent Scanner. |
| | Grid_Name | text | Grid template name or grid file name |
| | Grid_Date | integer | Date the grid template or grid file was created |
| | Grid_NumSubGridRows | integer | Number of subgrid columns |
| | Grid_NumSubGridCols | integer | Number of subgrid columns |
| | Grid_NumRows | integer | Number of spots per row of each subgrid |
| | Grid_NumCols | integer | Number of spots per column of each subgrid |
| | Grid_RowSpacing | float | Space between rows on the grid |
| | Grid_ColSpacing | float | Space between column on the grid |
| | Grid_OffsetX | float | In a dense pack array, the offset in the X direction |

**Table 4** List of parameters and options contained within the COMPACT text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Grid_OffsetY | float | In a dense pack array, the offset in the Y direction |
| | Grid_NomSpotWidth | float | Nominal width in microns of a spot from grid |
| | Grid_NomSpotHeight | float | Nominal height in microns of a spot from grid |
| | Grid_GenomicBuild | text | The build of the genome used to create the annotation (if available). If the genome build is not available (not all designs have this information), then it is not put out. All recent and all future designs have it. |
| | FeatureExtractor_Barcode | text | Barcode of the Agilent microarray read from the scan image |
| | FeatureExtractor_Sample | text | Names of hybridized samples (red/green) |
| | FeatureExtractor_ScanFileName | text | Name of the scan file used for Feature Extraction |
| | FeatureExtractor_ArrayName | text | Microarray filename |
| | FeatureExtractor_ScanFileGUID | text | GUID of the scan file |
| | FeatureExtractor_DesignFileName | text | Design or grid file used for Feature Extraction |
| | FeatureExtractor_ExtractionTime | text | Time stamp at the beginning of Feature Extraction |
| | FeatureExtractor_UserName | text | Windows Log-In Name of the User who ran Feature Extraction |
| | FeatureExtractor_ComputerName | text | Computer name on which Feature Extraction was run |
| | FeatureExtractor_Version | text | Version of Feature Extractor |
| | FeatureExtractor_IsXDRExtraction | integer<br>1 = True<br>0 = False | Says if result is from an XDR extraction |

**Table 4**    List of parameters and options contained within the COMPACT text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | FeatureExtractor_ColorMode | integer | A flag to indicate output color |
| | | 0 | One color; green only |
| | | 1 | 2-color |
| | FeatureExtractor_QCReportType | integer | Type of QC report to generate |
| | | 0 | Gene Expression |
| | | 1 | CGH_ChIP |
| | | 2 | miRNA |
| | | 4 | Streamlined CGH |
| | DyeNorm_NormFilename | text | Name of the dye normalization list file |
| | DyeNorm_NormNumProbes | integer | Number of probes in the dye normalization list |
| | Grid_IsGridFile | boolean | |

# QC FEPARAMS Table

**Table 5** List of parameters and options contained within the QC text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Protocol _Name | text | Name of protocol used |
| | Protocol_date | text | Date the protocol was last modified |
| | Scan_ScannerName | text | Agilent scanner serial number used |
| | Scan_NumChannels | integer | Number of channels in the scan image |
| | Scan_date | text | Date the image was scanned |
| | Scan_MicronsPerPixelX | float | Number of microns per pixel in the X axis of the scan image |
| | Scan_MicronsPerPixelY | float | Number of microns per pixel in the Y axis of the scan image |
| | Scan_OriginalGUID | text | The global unique identifier for the scan image |
| | Scan_NumScanPass | 1 or 2 | For 5 micron scans, indicates whether the scan mode was a single (1) or double-pass scan mode on the Agilent Scanner. |
| | Grid_Name | text | Grid template name or grid file name |
| | Grid_Date | integer | Date the grid template or grid file was created |
| | Grid_NumSubGridRows | integer | Number of subgrid columns |
| | Grid_NumSubGridCols | integer | Number of subgrid columns |
| | Grid_NumRows | integer | Number of spots per row of each subgrid |
| | Grid_NumCols | integer | Number of spots per column of each subgrid |
| | Grid_RowSpacing | float | Space between rows on the grid |
| | Grid_ColSpacing | float | Space between column on the grid |
| | Grid_OffsetX | float | In a dense pack array, the offset in the X direction |

**Table 5**    List of parameters and options contained within the QC text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Grid_OffsetY | float | In a dense pack array, the offset in the Y direction |
| | Grid_NomSpotWidth | float | Nominal width in microns of a spot from grid |
| | Grid_NomSpotHeight | float | Nominal height in microns of a spot from grid |
| | Grid_GenomicBuild | text | The build of the genome used to create the annotation (if available). If the genome build is not available (not all designs have this information), then it is not put out. All recent and all future designs have it. |
| | FeatureExtractor_Barcode | text | Barcode of the Agilent microarray read from the scan image |
| | FeatureExtractor_Sample | text | Names of hybridized samples (red/green) |
| | FeatureExtractor_ScanFileName | text | Name of the scan file used for Feature Extraction |
| | FeatureExtractor_ArrayName | text | Microarray filename |
| | FeatureExtractor_ScanFileGUID | text | GUID of the scan file |
| | FeatureExtractor_DesignFileName | text | Design or grid file used for Feature Extraction |
| | FeatureExtractor_ExtractionTime | text | Time stamp at the beginning of Feature Extraction |
| | FeatureExtractor_UserName | text | Windows Log-In Name of the User who ran Feature Extraction |
| | FeatureExtractor_ComputerName | text | Computer name on which Feature Extraction was run |
| | FeatureExtractor_Version | text | Version of Feature Extractor |
| | FeatureExtractor_IsXDRExtraction | integer<br>1 = True<br>0 = False | Says if result is from an XDR extraction |

**Table 5** List of parameters and options contained within the QC text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | FeatureExtractor_ColorMode | integer | A flag to indicate output color |
| | | 0 | One color; green only |
| | | 1 | 2-color |
| | FeatureExtractor_QCReportType | integer | Type of QC report to generate |
| | | 0 | Gene Expression |
| | | 1 | CGH_ChIP |
| | | 2 | miRNA |
| | | 4 | Streamlined CGH |
| | DyeNorm_NormFilename | text | Name of the dye normalization list file |
| | DyeNorm_NormNumProbes | integer | Number of probes in the dye normalization list |
| | Grid_IsGridFile | boolean | Indicates whether the grid is from a grid file. |

# MINIMAL FEPARAMS Table

**Table 6**    List of parameters and options contained within the MINIMAL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Protocol _Name | text | Name of protocol used |
| | Protocol_date | text | Date the protocol was last modified |
| | Scan_ScannerName | text | Agilent scanner serial number used |
| | Scan_NumChannels | integer | Number of channels in the scan image |
| | Scan_date | text | Date the image was scanned |
| | Scan_MicronsPerPixelX | float | Number of microns per pixel in the X axis of the scan image |
| | Scan_MicronsPerPixelY | float | Number of microns per pixel in the Y axis of the scan image |
| | Scan_OriginalGUID | text | The global unique identifier for the scan image |
| | Scan_NumScanPass | 1 or 2 | For 5 micron scans, indicates whether the scan mode was a single (1) or double-pass scan mode on the Agilent Scanner. |
| | Grid_Name | text | Grid template name or grid file name |
| | Grid_Date | integer | Date the grid template or grid file was created |
| | Grid_NumSubGridRows | integer | Number of subgrid columns |
| | Grid_NumSubGridCols | integer | Number of subgrid columns |
| | Grid_NumRows | integer | Number of spots per row of each subgrid |
| | Grid_NumCols | integer | Number of spots per column of each subgrid |
| | Grid_RowSpacing | float | Space between rows on the grid |
| | Grid_ColSpacing | float | Space between column on the grid |
| | Grid_OffsetX | float | In a dense pack array, the offset in the X direction |

**Table 6** List of parameters and options contained within the MINIMAL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | Grid_OffsetY | float | In a dense pack array, the offset in the Y direction |
| | Grid_NomSpotWidth | float | Nominal width in microns of a spot from grid |
| | Grid_NomSpotHeight | float | Nominal height in microns of a spot from grid |
| | Grid_GenomicBuild | text | The build of the genome used to create the annotation (if available). If the genome build is not available (not all designs have this information), then it is not put out. All recent and all future designs have it. |
| | FeatureExtractor_Barcode | text | Barcode of the Agilent microarray read from the scan image |
| | FeatureExtractor_Sample | text | Names of hybridized samples (red/green) |
| | FeatureExtractor_ScanFileName | text | Name of the scan file used for Feature Extraction |
| | FeatureExtractor_ArrayName | text | Microarray filename |
| | FeatureExtractor_ScanFileGUID | text | GUID of the scan file |
| | FeatureExtractor_DesignFileName | text | Design or grid file used for Feature Extraction |
| | FeatureExtractor_ExtractionTime | text | Time stamp at the beginning of Feature Extraction |
| | FeatureExtractor_UserName | text | Windows Log-In Name of the User who ran Feature Extraction |
| | FeatureExtractor_ComputerName | text | Computer name on which Feature Extraction was run |
| | FeatureExtractor_Version | text | Version of Feature Extractor |
| | FeatureExtractor_IsXDRExtraction | integer<br>1 = True<br>0 = False | Says if result is from an XDR extraction |

**Table 6** List of parameters and options contained within the MINIMAL text output file (FEPARAMS table)

| Protocol Step | Parameters | Type/Options | Description |
|---|---|---|---|
| | FeatureExtractor_ColorMode | integer | A flag to indicate output color |
| | | 0 | One color; green only |
| | | 1 | 2-color |
| | FeatureExtractor_QCReportType | integer | Type of QC report to generate |
| | | 0 | Gene Expression |
| | | 1 | CGH_ChIP |
| | | 2 | miRNA |
| | | 4 | Streamlined CGH |
| | DyeNorm_NormFilename | text | Name of the dye normalization list file |
| | DyeNorm_NormNumProbes | integer | Number of probes in the dye normalization list |
| | Grid_IsGridFile | boolean | |

# Statistical results (STATS)

This middle section of the text file describes the results from the global array-wide statistical calculations. The STATS results are reported to 9 decimal places in exponential notation for all results files (FULL, COMPACT, QC, or MINIMAL).

## STATS Table (ALL text output types)

**Table 7**    Stats results contained in the text output file (STATS table)[*]

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
| --- | --- | --- | --- |
| gDarkOffsetAverage | rDarkOffsetAverage | float | Average dark offset per image per channel as measured by scanner |
| gDarkOffsetMedian | rDarkOffsetMedian | float | Median dark offset per image per channel as measured by the scanner |
| gDarkOffsetStdDev | rDarkOffsetStdDev | float | Standard deviation of the data points measured by the scanner to determine the dark offset per image per channel. |
| gDarkOffsetNumPts | rDarkOffsetNumPts | integer | Number of points of data measured by the scanner to determine the dark offset per image per channel |
| gSaturationValue | rSaturationValue | integer | Signal intensity at which spot is considered saturated. |
| gAvgSig2BkgeQC | rAvgSig2BkgeQC | float | The average ratio of net signal to local background for all spike-in probes |
| gAvgSig2BkgNegCtrl | rAvgSig2BkgNegCtrl | float | The average ratio of net signal to local background for all negative control probes |
| gRatioSig2BkgeQC_NegCtrl | rRatioSig2BkgeQC_NegCtrl | float | The ratio of AvgSig2BkgeQC to AvgSig2BkgNegCtrl |
| gNumSatFeat | rNumSatFeat | integer | The number of saturated features on the microarray per channel |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gLocalBGInlierNetAve | rLocalBGInlierNetAve | float | The average of the net signal of all inlier local backgrounds |
| gLocalBGInlierAve | rLocalBGInlierAve | float | The average of all inlier local backgrounds |
| gLocalBGInlierSDev | rLocalBGInlierSDev | float | The standard deviation of all inlier local backgrounds |
| gLocalBGInlierNum | rLocalBGInlierNum | integer | The number of inlier local backgrounds |
| gGlobalBGInlierAve | rGlobalBGInlierAve | float | The average of all inliers used in background estimation for the selected global background subtraction method or the average of all inlier local backgrounds if the local background subtraction method is selected (after global background adjustment is applied, if selected) |
| gGlobalBGInlierSDev | rGlobalBGInlierSDev | float | The standard deviation of all inliers used in background estimation for the selected global background subtraction method or the standard deviation of all inlier local backgrounds if the local background subtraction method is selected |
| gGlobalBGInlierNum | rGlobalBGInlierNum | integer | The number of all inliers used in background estimation for the selected global background subtraction method or the number of all inlier local backgrounds if the local background subtraction method is selected |
| gNumFeatureNonUnifOL | rNumFeatureNonUnifOL | integer | The number of features that are flagged as non-uniformity outliers |
| gNumPopnOL | rNumPopnOL | integer | The number of features that are flagged as population outliers |
| gNumNonUnifBGOL | rNumNonUnifBGOL | integer | The number of local background regions that are flagged as non-uniformity outliers |
| gNumPopnBGOL | rNumPopnBGOL | integer | The number of local background regions that are flagged as population outliers |
| gOffsetUsed | rOffsetUsed | float | Software estimated scanner offset |

**Table 7**    Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gGlobalFeatInlierAve | rGlobalFeatInlierAve | float | Average of all inlier features |
| gGlobalFeatInlierSDev | rGlobalFeatInlierSDev | float | Standard deviation of all inlier features |
| gGlobalFeatInlierNum | rGlobalFeatInlierNum | float | Number of all inlier features |
| AllColorPrcntSat | | float | The percentage of features that are saturated in both the green AND red channels |
| AnyColorPrcntSat | | float | The percentage of features that are saturated in either the green or red channel |
| AnyColorPrcntFeatNonUnifOL | | float | The percentage of features that are feature non-uniformity outliers in either channel |
| AnyColorPrcntBGNonUnifOL | | float | The percentage of local backgrounds that are non-uniformity outliers in either channel |
| AnyColorPrcntFeatPopnOL | | float | The percentage of features that are population outliers in either the green or red channel |
| AnyColorPrcntBGPopnOL | | float | The percentage of local backgrounds that are population outliers in either channel |
| TotalPrcntFeatOL | | float | The percentage of non-control features that are feature non-uniformity outliers in either the green or red channel or are saturated in both channels |
| gBGAdjust | rBGAdjust | float | Background offset constant to adjust all feature signals. If Adjust Background Globally is set True, all feature signals are adjusted by this offset. If set to the value entered in the protocol, all feature signals are adjusted so that very low level feature signals equal the protocol value. |
| gNumNegBGSubFeat | rNumNegBGSubFeat | integer | Number of background-subtracted features with negative signals |

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gNonCtrlNumNegFeatBGSubSig | rNonCtrlNumNegFeatBGSubSig | integer | Number of non-control features with negative background-subtracted signals |
| gLinearDyeNormFactor | rLinearDyeNormFactor | float | Global dye norm factor |
| gRMSLowessDNF | rRMSLowessDNF | float | The root mean square of the average lowess dye norm factor. The lowess dye norm factor for each feature is its DyeNormSignal divided by its BGSubSignal. |
| DyeNormDimensionlessRMS | | float | Dimensionless RMS correction metric (metric that indicates how much correction has been applied based upon the LOWESS curve) |
| DyeNormUnitWeightedRMS | | float | Unit weighted RMS correction metric (metric that indicates how much correction has been applied based upon the LOWESS curve) |
| gSpatialDetrendRMSFit | rSpatialDetrendRMSFit | float | Root mean square (RMS) of the fitted data points obtained from the Loess algorithm. This gives an idea of the curvature of the surface fit. |
| gSpatialDetrendRMS Filtered MinusFit | rSpatialDetrendRMS Filtered MinusFit | float | Approximate residual from the surface fit. |
| gSpatialDetrendSurfaceArea | rSpatialDetrendSurfaceArea | float | Normalized area–the fitted surface area divided by the projected area on the microarray; also gives an idea of the curvature of the surface gradient. |
| gSpatialDetrendVolume | rSpatialDetrendVolume | float | Sum of the intensities of the surface area minus the offset. The offset is calculated as the volume under the flat surface (parallel to the glass slide) passing through the minimum intensity point of the fitted surface. This number (total volume - offset) is normalized by the area of the microarray. |
| gSpatialDetrendAveFit | rSpatialDetrendAveFit | float | Describes the average intensity of the surface gradient |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gNonCtrlNumSatFeat | rNonCtrlNumSatFeat | integer | The number of saturated non-control features |
| gNonCtrl99PrcntNetSig | rNonCtrl99PrcntNetSig | float | NetSignal intensity at 99th percentile for all non-control probes |
| gNonCtrl50PrcntNetSig | rNonCtrl50PrcntNetSig | float | NetSignal intensity at 50th percentile for all non-control probes |
| gNonCtrl1PrcntNetSig | rNonCtrl1PrcntNetSig | float | NetSignal intensity at 1st percentile for all non-control probes |
| gNonCtrlMedPrcntCVBGSubSig | rNonCtrlMedPrcntCVBGSubSig | float | The median percent CV of background-subtracted signals for inlier noncontrol probes |
| gCtrleQCNumSatFeat | rCtrleQCNumSatFeat | integer | The number of saturated spike-in features |
| gCtrleQC99PrcntNetSig | rCtrleQC99PrcntNetSig | float | NetSignal intensity at 99th percentile of all spike-in probes |
| gCtrleQC50PrcntNetSig | rCtrleQC50PrcntNetSig | float | NetSignal intensity at 50th percentile of all spike-in probes |
| gCtrleQC1PrcntNetSig | rCtrleQC1PrcntNetSig | float | NetSignal intensity at 1st percentile of all spike-in probes |
| geQCMedPrcntCVBGSubSig | reQCMedPrcntCVBGSubSig | float | The median percent CV of background-subtracted signals for inlier spike-in probes |
| geQCSig2BkgLow1 | reQCSig2BkgLow1 | float | Median ratio (net signal to BGUsed) of all inlier features for an spike-in probe with lowest concentration spiked in red and green channels |
| geQCSig2BkgLow2 | reQCSig2BkgLow2 | float | Median ratio (net signal to BGUsed) of all inlier features for an spike-in probe with second lowest concentration spiked in red and green channels |
| gNegCtrlNumInliers | rNegCtrlNumInliers | integer | Number of all inlier negative controls |
| gNegCtrlAveNetSig | rNegCtrlAveNetSig | float | Average net signal of all inlier negative controls |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gNegCtrlSDevNetSig | rNegCtrlSDevNetSig | float | Standard deviation of the net signal of all inlier negative controls |
| gNegCtrlAveBGSubSig | rNegCtrlAveBGSubSig | float | Average background-subtracted signal of all inlier negative controls |
| gNegCtrlSDevBGSubSig | rNegCtrlSDevBGSubSig | float | Standard deviation of the background-subtracted signals of all inlier negative controls |
| gAveNumPixOLLo | rAveNumPixOLLo | integer | The average number of pixels that are rejected from each feature at the low end of the intensity spectrum |
| gAveNumPixOLHi | rAveNumPixOLHi | integer | The average number of pixels that are rejected from each feature at the high end of the intensity spectrum |
| gPixCVofHighSignalFeat | rPixCVofHighSignalFeat | float | Average of pixel CV for features with high signal |
| gNumHighSignalFeat | rNumHighSignalFeat | integer | The number of features with high signal |
| NonCtrlAbsAveLogRatio | | float | This result is from a two-step calculation. Step 1 for each probe calculates the absolute average log ratio of all inlier non-control features with minimum number of replicates. Step 2 calculates the average of all absolute average log ratios calculated in step 1. |
| NonCtrlSDevLogRatio | | float | The average standard deviation of log ratios of all inlier non-control probe sets with a minimum number of replicates |
| NonCtrlSNRLogRatio | | float | The average of signal to noise values of the log ratio for all inlier non-control probe sets with a minimum number of replicates |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| eQCAbsAveLogRatio | | float | This result is from a two-step calculation. Step 1 for each probe calculates the absolute average log ratio of all inlier spikein features with minimum number of replicates. Step 2 calculates the average of all absolute average log ratios calculated in step 1. |
| eQCSDevLogRatio | | float | Average standard deviation of log ratios of all inlier spike-in probe sets with a minimum number of replicates |
| eQCSNRLogRatio | | float | Average signal to noise value of log ratios of all inlier spike-in probe sets with a minimum number of replicates |
| AddErrorEstimateGreen | | float | The additive error estimated for the microarray in the green channel. |
| AddErrorEstimateRed | | float | The additive error estimated for the microarray in the red channel. |
| TotalNumFeatures | | integer | Total number of features that show up in output file. |
| NonCtrlNumUpReg | | integer | Number of up-regulated non-control probes |
| NonCtrlNumDownReg | | integer | Number of down-regulated non-control probes |
| eQCObsVsExpLRSlope | | float | For 2-color QC report: Slope of the linear regression fit of the plot of the expected versus observed average log ratio for each spike-in probe |
| eQCObsVsExpLRIntercept | | float | For 2-color QC report: Intercept of the linear regression fit of the plot of the expected versus observed average log ratio for each spike-in probe |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| eQCObsVsExpCorr | | float | For 2-color QC report: The R2 value of the linear regression fit of the plot of the expected versus observed average log ratio for each spike-in probe |
| NumIsNorm | | integer | Number of features used for normalization |
| ROI Width ROI Height | | float | The width or height (in pixels) of the region of interest (ROI) about a nominal spot location. The spotfinder determines the found centroid and spot size of the spot within the ROI. |
| CentroidDiffX | | float | The average absolute of difference between nominal centroids and corresponding found centroids in X direction |
| CentroidDiffY | | float | The average absolute of difference between nominal centroids and corresponding found centroids in Y direction |
| NumFoundFeat | | integer | The number of features that are flagged as found |
| MaxNonUnifEdges | | float | Maximum fraction of features that are non-uniform along any edge of the microarray |
| MaxSpotNotFoundEdges | | float | Maximum fraction of features that are not found along any edge of the microarray |
| gMultDetrendRMS Fit | rMultDetrendRMS Fit | float | Root mean square (RMS) of the fitted data points obtained from the second degree polynomial equation in Multiplicative Detrending. This gives an idea of the curvature of the surface fit to the "hybridization dome" in the Agilent Hybridization chambers. |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gMultDetrendSurfaceAverage | rMultDetrendSurfaceAverage | float | The average of the surface calculated by multiplicative detrending. This average is used to normalize the surface. It is a straight average over all the points in the surface. |
| DerivativeOfLogRatioSD | | float | Measures the standard deviation of the probe-to-probe difference of the log ratios. This is a metric used in CGH experiments where differences in the log ratios are small on average. A smaller standard deviation here indicates less noise in the biological signals. |
| eQCLowSigName1 | | text | The probe name of the eQC probe spiked in at the lowest concentration. |
| eQCLowSigName2 | | text | The probe name of the eQC probe spiked in at the second lowest concentration. |
| eQCOneColorLogLowSignal | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Log of low signal for the data |
| eQCOneColorLogLowSignal-Error | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Error in the log of low signal for the data |
| eQCOneColorLogHighSignal | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Log of high signal for the data |
| eQCOneColorLinFitLogLowConc | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Log of low concentration in the linear range of curve fit |
| eQCOneColorLinFitLogLow-Signal | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Log of low signal in the linear range of curve fit |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| eQCOneColorLinFitLogHigh-Conc | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Log of high concentration in the linear range of curve fit |
| eQCOneColorLinFitLogHigh-Signal | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Log of high signal in the linear range of curve fit |
| eQCOneColorLinFitSlope | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Slope of the linear range of curve fit |
| eQCOneColorLinFitIntercept | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Intercept of the linear range of curve fit |
| eQCOneColorLinFitRSQ | | float | Agilent Spike-In Concentration-Response Statistic in the 1-color QC Report: Square of the correlation coefficient of the linear range of curve fit. |
| eQCOneColorSpikeDetection-Limit | | float | The detection limit as determined by measuring the average plus 1 standard deviation of all spike-in probes below the linear concentration range. This value is the maximum of these. |
| gNonCtrl50PrcntBGSubSig | gNonCtrl50PrcntBGSubSig | float | Background-subtracted signal intensity at 50th percentile for all non-control probes. |
| gCtrleQC50PrcntBGSubSig | rCtrleQC50PrcntBGSubSig | float | The median background-subtracted signal for all the embedded QC probes on the microarray. |

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gMedPrcntCVProcSignal | rMedPrcntCVProcSignal | float | The median %CV for replicate non-control probes using the processed signal. This value is calculated by calculating the average, SD and %CV of the processed signal of each replicated probe. For non-control replicated probes, there must be at least 10 CVs from which to calculate a median; otherwise, -1 is reported. The MedPrcntCVProcSignal and the MedPrcntCVBGSubSignal show if Multiplicative Detrending is having a positive effect on the data. If multiplicative detrending is helping, the MedPrcntCVProcSignal should be smaller than the MedPrcntCVBGSubSignal. |
| geQCMedPrcntCVProcSignal | reQCMedPrcntCVProcSignal | float | This is the same as MedPrcntCVProcSignal, except that it is performed using the eQC SpikeIn Replicates rather than the nonControl Replicates. There must be at least 3 CVs from which to calculate a median. |
| gOutlierFlagger_Auto_FeatB Term | rOutlierFlagger_Auto_FeatB Term | float | Applies to feature: specifies the variance due to the Poisson distributed noise; automatically calculated when OLAutoCompute is turned on |
| gOutlierFlagger_Auto_FeatC Term | rOutlierFlagger_Auto_FeatC Term | float | Applies to feature: specifies variance due to background noise of the scanner, slide glass, and other signal-independent sources; automatically calculated when OLAutoCompute is turned on |
| gOutlierFlagger_Auto_BgndB Term | rOutlierFlagger_Auto_BgndB Term | float | Applies to background: specifies the variance due to the Poisson distributed noise; automatically calculated when OLAutoCompute is turned on |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gOutlierFlagger_Auto_BgndC Term | rOutlierFlagger_Auto_BgndC Term | float | Applies to background: specifies variance due to background noise of the scanner, slide glass, and other signal-independent sources; automatically calculated when OLAutoCompute is turned on |
| OutlierFlagger_FeatChiSq | | float | Confidence Interval for the feature |
| OutlierFlagger_BgndChiSq | | float | Confidence Interval for the background |
| gXDRLowPMTSlope | rXDRLowPMTSlope | | The slope that is multiplied by the original low intensity Mean Signal to get the XDR mean signal. Used in the linear equation relating the Mean (or Median) Signal in the low intensity scan to the scaled intensity used in the combined XDR output. |
| gXDRLowPMTIntercept | rXDRLowPMTIntercept | | The intercept that is added to the Slope*LowIntensityMeanSignal to get the XDR Mean Signal. Used in the linear equation relating the Mean (or Median) Signal in the low intensity scan to the scaled intensity used in the combined XDR output. |
| GriddingStatus | | integer | Indicates that the automatic image processing was flagged as needing evaluation. |
| NumGeneNonUnifOL | | integer | Number of genes that do not have any replicate features on the array where both color channels are not Feature Non-Uniform outliers. If multiple probes address the same gene, this value actually states the number of probes that have no non-uniform replicates. |
| TotalNumberOfReplicated Genes | | integer | Number of genes that have replicate features on the array. |

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| EffectiveFeatureSizeFraction | | float | Estimates the ratio of the effective feature size to the nominal feature size. It is calculated by looking at the ratio of the whole spot measurement versus the cookie measurement. |
| Feature UniformityAnomaly Fraction | | float | Fraction (Num/TotalNum) of the number of features looked at that had anomalous ratios. This gives a measure of the percentage of representative spots that are strange (e.g., donuts, super hot spots, hot crescents). |
| UsedDefaultEffectiveFeature Size | | integer | Reports whether or not the default effective feature size was used. If the default was used, the stat is 1. If the effective feature size was estimated, the stat value is 0. |
| gPercentileIntensityProcessed Signal | rPercentileIntensityProcessed Signal | float | The protocol lets you enter the Percentile Value at which the intensity of the noncontrol signals is recorded. All protocols specify the 75th percentile. This number is the intensity of all the noncontrol signals in the 75th percentile. This stat is used to normalize 1-color data. |
| gTotalSignal99pctile | | float | These are metrics for miRNA only. This is the value of the TotalGeneSignal for all genes at the 99th percentile. |
| gNegCtrlSpread | rNegCtrlSpread | float | The root mean square (RMS) of the preliminary spatial fit of the negative controls. It is equivalent to a standard deviation of NC signals after removal of spatial homogeneities. Used as a preliminary estimation of the noise on the array for selecting near-zero probes in spatial detrending, and conversely for excluding near-zero probes in multiplicative detrending. |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| gNonCtrlNumWellAboveBG | rNonCtrlNumWellAboveBG | integer | Measure of the number of noncontrol features whose signals are well above background. Used as a metric for the number of features with significant signal. |
| LogRatioImbalance | | float | This metric is for CGH only. It calculates the amount of amplifications versus duplications per chromosome to determine if there is an imbalance that falls outside of normal expectations. |
| ImageDepth | | string | 16 bit or 20 bit |
| AFHold | | float | The percentage of time, during a scan that the Autofocus assembly holds its position rather than actively maintaining focus. Typically, the value is less than 2%; however, the value will be larger if there are obstructions on the microarray that interfere with the laser beams. |
| gPMTVolts | rPMTVolts | float | The voltages that Photomultipliers are set to. The voltage adjusts the spectral response of the scanner to incoming light from the lasers. In general, the higher the PMTVoltage, the higher the signals will be for fluorescent artifacts that are scanned. Typical numbers here are between 350 – 525 mV, but can vary depending on the PMT. |
| GlassThickness | | float | Expressed in microns. This represents the thickness of the microarray slide, as measured during autofocus homing. Using standard Agilent slides, the values range from 900 – 1000. Nominal values for non-Agilent slides are specified between 900 and 1100 for C scanners, and 900 and 1200 for B scanners. |

**Table 7**    Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| RestrictionControl | | float | Restriction control probes are a set of probes spanning cut sites that are not variant in samples. If the protocol is followed correctly, these probes should always give 0 signal. The final restriction control value is the minimum of the restriction control values of red channel and green channel. If restriction control probes are not present in the design, the RestrictionControl value is set to "-1". |
| GridHasBeenOptimized | | boolean 0 = False 1 = True | Indicates if grid has been adjusted for better fit as result of performing the interactively adjust corners method. |
| ExtractionStatus | | integer 0=in range; 1=out of range | This is put out only if a metric set has been run. It gives a status of the overall array. |
| QCMetricResults | | String | If the Extraction Status = 0, the output says ExtractionInRange. If the Extraction Status = 1, the output says ExtractionEvaluate. |
| UpRandomnessRatio | | float | Variance measure of whether or not positive Log Ratios appear to be correlated with position on the array |
| DownRandomnessRatio | | float | Variance measure of whether or not negative Log Ratios appear to be correlated with position on the array |
| UpRandomnessSDRatio | | float | StDev measure of whether or not positive Log Ratios appear to be correlated with position on the array |
| DownRandomnessSDRatio | | float | StDev measure of whether or not negative Log Ratios appear to be correlated with position on the array |

**Table 7** Stats results contained in the text output file (STATS table)[*] (continued)

| Stats (Green Channel) | Stats (Red Channel) | Type | Description |
|---|---|---|---|
| Metric_MetricName | | | (Optional. Only displayed when a metric set is used.) The name of a metric in the metric set. The given value is the one that has been calculated for this metric. You can have more than one metric in a given metric set. |
| Metric_MetricName_IsInRange | | integer 1=in range; 0=out of range | (Optional. Only displayed when a metric set is used.) Indicates whether the metric was within any user-defined thresholds found in the metric set for that metric. |

[*] Results are reported to 9 decimal places in exponential notation for all result files.

# Feature results (FEATURES)

The bottom section of the text file gives descriptions of the results for each feature. Results are reported to 9 decimal places in exponential notation for all result files.

## FULL Features Table

**Table 8**    Feature results contained in the FULL output text file (FULL FEATURES table)[*]

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| FeatureNum | | integer | | Feature number |
| Row | | integer | | Feature location: row |
| Col | | integer | | Feature location: column |
| Accessions | | text | | Gene accession numbers |
| Chr_coord | | text | | Chromosome coordinates of the feature |
| SubTypeMask | | integer | | Numeric code defining the subtype of any control feature |
| SubTypeName | | integer | | Name of the subtype of any control feature |
| Start | | integer | | Indicates the place in the transcript where the probe sequence starts. |
| Sequence | | text | | The sequence of bases printed on the array. |
| ProbeUID | | integer | | Unique integer for each unique probe in a design |

**Table 8** Feature results contained in the FULL output text file (FULL FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| ControlType | | integer | | Feature control type (See "XML Control Type output" on page 156 for definitions.) |
| | | | 0<br>1<br>-1<br>-15000<br>-20000<br>-30000 | Control type none<br>Positive control<br>Negative control<br>SNP<br>Not probe (See Ch. 4 for definition)<br>Ignore (See Ch. 4 for definition) |
| ProbeName | | text | | An Agilent-assigned identifier for the probe synthesized on the microarray |
| GeneName | | text | | This is an identifier for the gene for which the probe provides expression information. The target sequence identified by the systematic name is normally a representative or consensus sequence for the gene. |
| SystematicName | | text | | This is an identifier for the target sequence that the probe was designed to hybridize with. Where possible, a public database identifier is used (e.g., TAIR locus identifier for Arabidopsis). Systematic name is reported ONLY if Gene name and Systematic name are different. |
| Description | | text | | Description of gene |
| PositionX<br>PositionY | | float | | Found coordinates of the feature centroid in microns |

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| LogRatio (base 10) | | float | | per feature, log of (rProcessedSignal/gProcessedSignal) |
| | | | | If SURROGATES are turned off, then: |
| | | | -4 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig > 0.0 |
| | | | 4 | if DyeNormRedSig > 0.0 & DyeNormGreenSig <= 0.0 |
| | | | 0 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig <= 0.0 |
| LogRatioError | | float | | If SURROGATES are turned off, then: |
| | | | 1000 | if DyeNormRedSig <= 0.0 OR DyeNormGreenSig <= 0.0 |
| | | | | IF SURROGATES are turned on, then: |
| | | | | LogRatioError = error of the log ratio calculated according to the error model chosen |
| PValueLogRatio | | float | | Significance level of the LogRatio computed for a feature |
| gSurrogateUsed | rSurrogateUsed | float | Non-zero value 0 | The g(r) surrogate value used No surrogate value used |
| gIsFound | rIsFound | boolean | 1 = IsFound 0 = IsNotFound | A boolean used to flag found features. The flag is applied independently in each channel. |
| | | | | A feature is considered Found if two conditions are true: 1) the difference between the feature signal and the local background signal is more than 1.5 times the local background noise and 2) the spot diameter is at least 0.30 times the nominal spot diameter. |

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gProcessedSignal | rProcessedSignal | float | | The signal left after all the Feature Extraction processing steps have been completed. In the case of one color, ProcesssedSignal contains the Multiplicatively Detrended BackgroundSubtracted Signal if the detrending is selected and helps. If the detrending does not help, this column will contain the BackgroundSubtractedSignal. |
| gProcessedSigError | rProcessedSigError | float | | The universal or propagated error left after all the processing steps of Feature Extraction have been completed. In the case of one color, ProcessedSignalError has had the Error Model applied and will contain at least the larger of the universal (UEM) error or the propagated error.<br>If multiplicative detrending is performed, ProcessedSignalError contains the error propagated from detrending. This is done by dividing the error by the normalized MultDetrendSignal. |
| gNumPixOLHi | rNumPixOLHi | integer | | Number of outlier pixels per feature with intensity > upper threshold set via the pixel outlier rejection method. The number is computed independently in each channel. These pixels are omitted from all subsequent calculations. |
| gNumPixOLLo | rNumPixOLLo | integer | | Number of outlier pixels per feature with intensity < lower threshold set via the pixel outlier rejection method. The number is computed independently in each channel. These pixels are omitted from all subsequent calculations.<br><br>NOTE: The pixel outlier method is the ONLY step that removes data in Feature Extraction. |

**Table 8** Feature results contained in the FULL output text file (FULL FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gNumPix | rNumPix | integer | | Total number of pixels used to compute feature statistics; i.e. total number of inlier pixels/per spot; same in both channels |
| gMeanSignal | rMeanSignal | float | | Raw mean signal of feature from inlier pixels in green and/or red channel |
| gMedianSignal | rMedianSignal | float | | Raw median signal of feature from inlier pixels in green and/or red channel |
| gPixSDev | rPixSDev | float | | Standard deviation of all inlier pixels per feature; this is computed independently in each channel. |
| gPixNormIQR | rPixNormIQR | float | | The normalized Inter-quartile range of all of the inlier pixels per feature. The range is computed independently in each channel. |
| gBGNumPix | rBGNumPix | integer | | Total number of pixels used to compute local BG statistics per spot; i.e. total number of BG inlier pixels; same in both channels |
| gBGMeanSignal | rBGMeanSignal | float | | Mean local background signal (local to corresponding feature) computed per channel (inlier pixels) |
| gBGMedianSignal | rBGMedianSignal | float | | Median local background signal (local to corresponding feature) computed per channel (inlier pixels) |
| gBGPixSDev | rBGPixSDev | float | | Standard deviation of all inlier pixels per local BG of each feature, computed independently in each channel |
| gBGPixNormIQR | rBGPixNormIQR | float | | The normalized Inter-quartile range of all of the inlier pixels per local BG of each feature. The range is computed independently in each channel. |
| gNumSatPix | rNumSatPix | integer | | Total number of saturated pixels per feature, computed per channel |

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gIsSaturated | rIsSaturated | boolean | 1 = Saturated or 0 = Not saturated | Boolean flag indicating if a feature is saturated or not. A feature is saturated IF 50% of the pixels in a feature are above the saturation threshold. |
| gIsLowPMTScaled Up | rIsLowPMTScaled Up | boolean | 1 = Low 0 = High | Reports if the feature signal value is from the scaled-up low signal image or from the high signal image |
| PixCorrelation | | float | | Ratio of estimated feature covariance in RedGreen space to product of feature standard deviation in Red Green space<br><br>The covariance of two features measures their tendency to vary together, i.e., to co-vary. In this case, it is a cumulative quantitation of the tendency of pixels belonging to a particular feature in Red and Green spaces to co-vary. |
| BGPixCorrelation | | float | | The same concept as above but in case of background. |
| gIsFeatNonUnifOL | rIsFeatNonUnifOL | boolean | g(r)IsFeatNonUnifOL = 1 indicates Feature is a non-uniformity outlier in g(r) | Boolean flag indicating if a feature is a NonUniformity Outlier or not. A feature is non-uniform if the pixel noise of feature exceeds a threshold established for a "uniform" feature. |
| gIsBGNonUnifOL | rIsBGNonUnifOL | boolean | g(r)IsBGNonUnifOL = 1 indicates Local background is a non-uniformity outlier in g(r) | The same concept as above but for background. |

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gIsFeatPopnOL | rIsFeatPopnOL | boolean | g(r)IsFeatPopnOL = 1 indicates Feature is a population outlier in g(r) | Boolean flag indicating if a feature is a Population Outlier or not. Probes with replicate features on a microarray are examined using population statistics. A feature is a population outlier if its signal is less than a lower threshold or exceeds an upper threshold determined using a multiplier (1.42) times the interquartile range (i.e., IQR) of the population. |
| gIsBGPopnOL | rIsBGPopnOL | boolean | g(r)IsBGPopnOL = 1 indicates local background is a population outlier in g(r) | The same concept as above but for background |
| IsManualFlag | | boolean | | Boolean to flag features for downstream filtering in third party gene expression software. |
| gBGSubSignal | rBGSubSignal | float | g(r)BGSubSignal = g(r)MeanSignal - g(r)BGUsed | Background-subtracted signal. To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| gBGSubSigError | rBGSubSigError | float | | Propagated standard error as computed on net g(r) background-subtracted signal. For one color, the error model is applied to the background-subtracted signal. This will contain the larger of he universal (UEM) error or the propagated error. |
| BGSubSigCorrela-tion | | float | | Ratio of estimated background-subtracted feature signal covariance in RG space to product of background-subtracted feature standard deviation in RG space |

**Table 8** Feature results contained in the FULL output text file (FULL FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gIsPosAndSignif | rIsPosAndSignif | Boolean | g(r)isPosAndSignif = 1 indicates Feature is positive and significant above background | Boolean flag, established via a 2-sided t-test, indicates if the mean signal of a feature is greater than the corresponding background (selected by user) and if this difference is significant. To display variables used in the t-test, see Table 20 on page 190. |
| gPValFeatEqBG | rPValFeatEqBG | float | | pValue from t-test of significance between g(r)Mean signal and g(r) background (selected by user) |
| gNumBGUsed | rNumBGUsed | integer | | Number of local background regions or features used to calculate the background used for background subtraction on this feature. |
| gIsWellAboveBG | rIsWellAboveBG | Boolean | | Boolean flag indicating if a feature is WellAbove Background or not, feature passes g(r)IsPosAndSignif and additionally the g(r)BGSubSignal is greater than 2.6*g(r)BG_SD. You can change the multiplier 2.6. |
| gBGUsed | rBGUsed | float | g(r)BGSubSignal = g(r)MeanSignal - g(r)BGUsed | Background used to subtract from the MeanSignal; variable also used in t-test. To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| gBGSDUsed | rBGSDUsed | float | | Standard deviation of background used in g(r) channel; variable also used in t-test and surrogate algorithms. To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| IsNormalization | | boolean | 1 = Feature used; 0 = Feature not used | A boolean flag which indicates if a feature is used to measure dye bias |

**Table 8** Feature results contained in the FULL output text file (FULL FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gDyeNormSignal | rDyeNormSignal | float | | The dye-normalized signal in the indicated channel |
| gDyeNormError | rDyeNormError | float | | The standard error associated with the dye-normalized signal |
| DyeNormCorrelation | | float | | Dye-normalized red and green pixel correlation |
| ErrorModel | | | 0 = Propagated model chosen by you or by software 1 = Universal error model chosen by you or by software | Indicates the error model that you chose for Feature Extraction or that the software uses if you have chosen the "Most Conservative" option |
| xDev | | float | | A signal-to-noise parameter used to calculate pValue; calculated differently depending on error model chosen |
| gSpatialDetrendIsIn FilteredSet | rSpatialDetrendIsIn FilteredSet | boolean | 1 = Feature in filtered set 0 = Feature not in filtered set | Set to true for a given feature if it is part of the filtered set used to detrend the background. This feature is considered part of the locally weighted lowest x% of features as defined by the *DetrendLowPassPercentage*. |
| gSpatialDetrend SurfaceValue | rSpatialDetrend SurfaceValue | float | | Value of the smoothed surface calculated by the Spatial detrend algorithm |
| gIsLowEnoughAdd Detrend | rIsLowEnoughAdd Detrend | boolean | | These points are considered to be in the background for the purposes of spatial detrending and multiplicative detrending. If the Boolean value is true for a given point, it will be used in spatial detrending and not in multiplicative detrending (depends on parameters). |
| SpotExtentX | | float | | Diameter of the spot (X-axis) |
| SpotExtentY | | float | | Diameter of the spot (Y-axis) |

**Table 8** Feature results contained in the FULL output text file (FULL FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gNetSignal | rNetSignal | float | | MeanSignal minus DarkOffset |
| gMultDetrendSignal | rMultDetrendSignal | float | | A surface is fitted through the log of the background-subtracted signal to look for multiplicative gradients. A normalized version of that surface interpolated at each point of the microarray is stored in MultDetrendSignal. The surface is normalized by dividing each point by the overall average of the surface. That average is stored in MultDetrendSurfaceAverage as a statistic. 1-color only |
| gProcessed Background | rProcessed Background | float | | Indicates the Background signal that was selected to be used (Mean or Median). |
| gProcessedBkng Error | rProcessedBkng Error | float | | Indicates the Background error that was selected to be used (PixSD or NormIQR) |
| IsUsedBGAdjust | | boolean | 1 = Feature used 0 = Feature not used | A Boolean used to flag features used for computation of global BG offset |
| gInterpolatedNeg CtrlSub | rInterpolatedNeg CtrlSub | float | | Value at the polynomial fit of the negative controls. |
| gIsInNegCtrlRange | rIsInNegCtrlRange | boolean | | Set to true for a given feature if its signal intensity is in the negative control range. |
| gIsUsedInMD | rIsUsedInMD | boolean | | Indicates whether this feature was included in the set used to generate the multiplicative detrend surface. |

[*] Results are reported to 9 decimal places in exponential notation for all result files.

# COMPACT Features Table

**Table 9** Feature results contained in the COMPACT output text file (COMPACT FEATURES table)[*]

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| FeatureNum | | integer | | Feature number |
| Row | | integer | | Feature location: row |
| Col | | integer | | Feature location: column |
| SubTypeMask | | integer | | Numeric code defining the subtype of any control feature |
| ControlType | | integer | | Feature control type (See "XML Control Type output" on page 156 for definitions.) |
| | | | 0 | Control type none |
| | | | 1 | Positive control |
| | | | -1 | Negative control |
| | | | -15000 | SNP |
| | | | -20000 | Not probe (See Ch. 4 for definition) |
| | | | -30000 | Ignore (See Ch. 4 for definition) |
| ProbeName | | text | | An Agilent-assigned identifier for the probe synthesized on the microarray |
| SystematicName | | text | | This is an identifier for the target sequence that the probe was designed to hybridize with. Where possible, a public database identifier is used (e.g., TAIR locus identifier for Arabidopsis). Systematic name is reported ONLY if Gene name and Systematic name are different. |
| Position X Position Y | | float | | Found coordinates of the feature centroid in microns |

**Table 9** Feature results contained in the COMPACT output text file (COMPACT FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| LogRatio (base 10) | | float | | per feature, log of (rProcessedSignal/gProcessedSignal) |
| | | | | If SURROGATES are turned off, then: |
| | | | -4 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig > 0.0 |
| | | | 4 | if DyeNormRedSig > 0.0 & DyeNormGreenSig <= 0.0 |
| | | | 0 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig <= 0.0 |
| LogRatioError | | float | | If SURROGATES are turned off, then: |
| | | | 1000 | if DyeNormRedSig <= 0.0 OR DyeNormGreenSig <= 0.0 |
| | | | | IF SURROGATES are turned on, then: |
| | | | | LogRatioError = error of the log ratio calculated according to the error model chosen |
| PValueLogRatio | | float | | Significance level of the Log Ratio computed for a feature |
| gProcessedSignal | rProcessedSignal | float | | The signal left after all the Feature Extraction processing steps have been completed. In the case of one color, ProcesssedSignal contains the Multiplicatively Detrended BackgroundSubtracted Signal if the detrending is selected and helps. If the detrending does not help, this column will contain the BackgroundSubtractedSignal. |

**Table 9** Feature results contained in the COMPACT output text file (COMPACT FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gProcessedSigError | rProcessedSigError | float | | The universal or propagated error left after all the processing steps of Feature Extraction have been completed. In the case of one color, ProcessedSignalError has had the Error Model applied and will contain at least the larger of the universal (UEM) error or the propagated error. <br> If multiplicative detrending is performed, ProcessedSignalError contains the error propagated from detrending. This is done by dividing the error by the normalized MultDetrendSignal. |
| gMedianSignal | rMedianSignal | float | | Raw median signal of feature in green (red) channel (inlier pixels) |
| gBGMedianSignal | rBGMedianSignal | float | | Median local background signal (local to corresponding feature) computed per channel (inlier pixels) |
| gBGPixSDev | rBGPixSDev | float | | Standard deviation of all inlier pixels per local BG of each feature, computed independently in each channel |
| gIsSaturated | rIsSaturated | boolean | 1 = Saturated or 0 = Not saturated | Boolean flag indicating if a feature is saturated or not. A feature is saturated IF 50% of the pixels in a feature are above the saturation threshold. |
| gIsLowPMTScaled Up | rIsLowPMTScaled Up | boolean | 1 = Low 0 = High | Reports if the feature signal value is from the scaled-up low signal image or from the high-signal image |
| gIsFeatNonUnifOL | rIsFeatNonUnifOL | boolean | g(r)IsFeatNonUnifO L = 1 indicates Feature is a non-uniformity outlier in g(r) | Boolean flag indicating if a feature is a NonUniformity Outlier or not. A feature is non-uniform if the pixel noise of feature exceeds a threshold established for a "uniform" feature. |

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gIsBGNonUnifOL | rIsBGNonUnifOL | boolean | g(r)IsBGNonUnifOL = 1 indicates Local background is a non-uniformity outlier in g(r) | The same concept as above but for background. |
| gIsFeatPopnOL | rIsFeatPopnOL | boolean | g(r)IsFeatPopnOL = 1 indicates Feature is a population outlier in g(r) | Boolean flag indicating if a feature is a Population Outlier or not. Probes with replicate features on a microarray are examined using population statistics.<br><br>A feature is a population outlier if its signal is less than a lower threshold or exceeds an upper threshold determined using a multiplier (1.42) times the interquartile range (i.e., IQR) of the population. |
| gIsBGPopnOL | rIsBGPopnOL | boolean | g(r)IsBGPopnOL = 1 indicates local background is a population outlier in g(r) | The same concept as above but for background |
| IsManualFlag | | boolean | | Flags features for downstream filtering in third party gene expression software. |
| gBGSubSignal | rBGSubSignal | float | g(r)BGSubSignal = g(r)MeanSignal - g(r)BGUsed | Background-subtracted signal. To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| gIsPosAndSignif | rIsPosAndSignif | boolean | g(r)isPosAndSignif = 1 indicates Feature is positive and significant above background | Boolean flag, established via a 2-sided t-test, indicates if the mean signal of a feature is greater than the corresponding background (selected by user) and if this difference is significant. To display variables used in the t-test, see Table 20 on page 190. |

**Table 9** Feature results contained in the COMPACT output text file (COMPACT FEATURES table)[*] (continued)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gIsWellAboveBG | rIsWellAboveBG | boolean | | Boolean flag indicating if a feature is WellAbove Background or not, feature passes g(r)IsPosAndSignif and additionally the g(r)BGSubSignal is greater than 2.6*g(r)BG_SD. You can change the multiplier 2.6. |
| SpotExtentX | | float | | Diameter of the spot (X-axis) |
| gBGMeanSignal | rBGMeanSignal | float | | Mean local background signal (local to corresponding feature) computed per channel (inlier pixels) |

[*] Results are reported to 9 decimal places in exponential notation for all result files.

# QC Features Table

**Table 10**  Feature results contained in the QC output text file (QC FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| FeatureNum | | integer | | Feature number |
| Row | | integer | | Feature location: row |
| Col | | integer | | Feature location: column |
| SubTypeMask | | integer | | Numeric code defining the subtype of any control feature |
| ControlType | | integer | | Feature control type (See "XML Control Type output" on page 156 for definitions.) |
| | | | 0<br>1<br>-1<br>-15000<br>-20000<br>-30000 | Control type none<br>Positive control<br>Negative control<br>SNP<br>Not probe (See Ch. 4 for definition)<br>Ignore (See Ch. 4 for definition) |
| ProbeName | | text | | An Agilent-assigned identifier for the probe synthesized on the microarray |
| SystematicName | | text | | This is an identifier for the target sequence that the probe was designed to hybridize with. Where possible, a public database identifier is used (e.g., TAIR locus identifier for Arabidopsis). Systematic name is reported ONLY if Gene name and Systematic name are different. |
| Description | | text | | Description of gene |
| PositionX<br>PositionY | | float | | Found coordinates of the feature centroid in microns |

**Table 10** Feature results contained in the QC output text file (QC FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| LogRatio (base 10) | | float | | per feature, log of (rProcessedSignal/gProcessedSignal) |
| | | | | If SURROGATES are turned off, then: |
| | | | -4 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig > 0.0 |
| | | | 4 | if DyeNormRedSig > 0.0 & DyeNormGreenSig <= 0.0 |
| | | | 0 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig <= 0.0 |
| LogRatioError | | float | | If SURROGATES are turned off, then: |
| | | | 1000 | if DyeNormRedSig <= 0.0 OR DyeNormGreenSig <= 0.0 |
| | | | | IF SURROGATES are turned on, then: |
| | | | | LogRatioError = error of the log ratio calculated according to the error model chosen |
| PValueLogRatio | | float | | Significance level of the LogRatio computed for a feature |
| gProcessedSignal | rProcessedSignal | float | | The signal left after all the Feature Extraction processing steps have been completed. In the case of one color, ProcesssedSignal contains the Multiplicatively Detrended BackgroundSubtracted Signal if the detrending is selected and helps. If the detrending does not help, this column will contain the BackgroundSubtractedSignal. |

**Table 10** Feature results contained in the QC output text file (QC FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gProcessedSigError | rProcessedSigError | float | | The universal or propagated error left after all the processing steps of Feature Extraction have been completed. In the case of one color, ProcessedSignalError has had the Error Model applied and will contain at least the larger of the universal (UEM) error or the propagated error.<br>If multiplicative detrending is performed, ProcessedSignalError contains the error propagated from detrending. This is done by dividing the error by the normalized MultDetrendSignal. |
| gNumPixOLHi | rNumPixOLHi | integer | | Number of outlier pixels per feature with intensity > upper threshold set via the pixel outlier rejection method. The number is computed independently in each channel. These pixels are omitted from all subsequent calculations. |
| gNumPixOLLo | rNumPixOLLo | integer | | Number of outlier pixels per feature with intensity < lower threshold set via the pixel outlier rejection method. The number is computed independently in each channel. These pixels are omitted from all subsequent calculations.<br><br>NOTE: The pixel outlier method is the ONLY step that removes data in Feature Extraction. |
| gNumPix | rNumPix | integer | | Total number of pixels used to compute feature statistics; i.e. total number of inlier pixels/per spot; same in both channels |
| gMeanSignal | rMeanSignal | float | | Raw mean signal of feature from inlier pixels in green and/or red channel |
| gMedianSignal | rMedianSignal | float | | Raw median signal of feature from inlier pixels in green and/or red channel |

**Table 10** Feature results contained in the QC output text file (QC FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gPixSDev | rPixSDev | float | | Standard deviation of all inlier pixels per feature; this is computed independently in each channel. |
| gBGMeanSignal | rBGMeanSignal | float | | Mean local background signal (local to corresponding feature) computed per channel (inlier pixels) |
| gBGMedianSignal | rBGMedianSignal | float | | Median local background signal (local to corresponding feature) computed per channel (inlier pixels) |
| gBGPixSDev | rBGPixSDev | float | | Standard deviation of all inlier pixels per local BG of each feature, computed independently in each channel |
| gIsSaturated | rIsSaturated | boolean | 1 = Saturated or 0 = Not saturated | Boolean flag indicating if a feature is saturated or not. A feature is saturated IF 50% of the pixels in a feature are above the saturation threshold. |
| gIsLowPMTScaled Up | rIsLowPMTScaled Up | boolean | 1 = Low 0 = High | Reports if the feature signal value is from the scaled-up low signal image or from the high signal image |
| BGPixCorrelation | | float | | The same concept as above but in case of background. |
| gIsFeatNonUnifOL | rIsFeatNonUnifOL | boolean | g(r)IsFeatNonUnifOL = 1 indicates Feature is a non-uniformity outlier in g(r) | Boolean flag indicating if a feature is a NonUniformity Outlier or not. A feature is non-uniform if the pixel noise of feature exceeds a threshold established for a "uniform" feature. |
| gIsBGNonUnifOL | rIsBGNonUnifOL | boolean | g(r)IsBGNonUnifOL = 1 indicates Local background is a non-uniformity outlier in g(r) | The same concept as above but for background. |

**Table 10** Feature results contained in the QC output text file (QC FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gIsFeatPopnOL | rIsFeatPopnOL | boolean | g(r)IsFeatPopnOL = 1 indicates Feature is a population outlier in g(r) | Boolean flag indicating if a feature is a Population Outlier or not. Probes with replicate features on a microarray are examined using population statistics. A feature is a population outlier if its signal is less than a lower threshold or exceeds an upper threshold determined using a multiplier (1.42) times the interquartile range (i.e., IQR) of the population. |
| gIsBGPopnOL | rIsBGPopnOL | boolean | g(r)IsBGPopnOL = 1 indicates local background is a population outlier in g(r) | The same concept as above but for background |
| IsManualFlag | | boolean | | Flags features for downstream filtering in third party gene expression software. |
| gBGSubSignal | rBGSubSignal | float | g(r)BGSubSignal = g(r)MeanSignal - g(r)BGUsed | Background-subtracted signal. To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| gIsPosAndSignif | rIsPosAndSignif | Boolean | g(r)isPosAndSignif = 1 indicates Feature is positive and significant above background | Boolean flag, established via a 2-sided t-test, indicates if the mean signal of a feature is greater than the corresponding background (selected by user) and if this difference is significant. To display variables used in the t-test, see Table 20 on page 190. |
| gIsWellAboveBG | rIsWellAboveBG | Boolean | | Boolean flag indicating if a feature is WellAbove Background or not, feature passes g(r)IsPosAndSignif and additionally the g(r)BGSubSignal is greater than $2.6*g(r)BG\_SD$. You can change the multiplier 2.6. |

**Table 10** Feature results contained in the QC output text file (QC FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| SpotExtentX | | float | | Diameter of the spot (X-axis) |
| gBGMeanSignal | rBGMeanSignal | float | | Mean local background signal (local to corresponding feature) computed per channel (inlier pixels) |

# MINIMAL Features Table

**Table 11**  Feature results contained in the MINIMAL output text file (MINIMAL FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| FeatureNum | | integer | | Feature number |
| Row | | integer | | Feature location: row |
| Col | | integer | | Feature location: column |
| ControlType | | integer | | Feature control type (See "XML Control Type output" on page 156 for definitions.) |
| | | | 0 | Control type none |
| | | | 1 | Positive control |
| | | | -1 | Negative control |
| | | | -15000 | SNP |
| | | | -20000 | Not probe (See Ch. 4 for definition) |
| | | | -30000 | Ignore (See Ch. 4 for definition) |
| ProbeName | | text | | An Agilent-assigned identifier for the probe synthesized on the microarray |
| SystematicName | | text | | This is an identifier for the target sequence that the probe was designed to hybridize with. Where possible, a public database identifier is used (e.g., TAIR locus identifier for Arabidopsis). Systematic name is reported ONLY if Gene name and Systematic name are different. |

**Table 11**  Feature results contained in the MINIMAL output text file (MINIMAL FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| LogRatio (base 10) | | float | | per feature, log of (rProcessedSignal/gProcessedSignal) |
| | | | | If SURROGATES are turned off, then: |
| | | | -4 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig > 0.0 |
| | | | 4 | if DyeNormRedSig > 0.0 & DyeNormGreenSig <= 0.0 |
| | | | 0 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig <= 0.0 |
| LogRatioError | | float | | If SURROGATES are turned off, then: |
| | | | 1000 | if DyeNormRedSig <= 0.0 OR DyeNormGreenSig <= 0.0 |
| | | | | IF SURROGATES are turned on, then: |
| | | | | LogRatioError = error of the log ratio calculated according to the error model chosen |
| PValueLogRatio | | float | | Significance level of the LogRatio computed for a feature |
| gProcessedSignal | rProcessedSignal | float | | The signal left after all the Feature Extraction processing steps have been completed. In the case of one color, ProcesssedSignal contains the Multiplicatively Detrended BackgroundSubtracted Signal if the detrending is selected and helps. If the detrending does not help, this column will contain the BackgroundSubtractedSignal. |

**Table 11** Feature results contained in the MINIMAL output text file (MINIMAL FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gProcessedSigError | rProcessedSigError | float | | The universal or propagated error left after all the processing steps of Feature Extraction have been completed. In the case of one color, ProcessedSignalError has had the Error Model applied and will contain at least the larger of the universal (UEM) error or the propagated error.<br>If multiplicative detrending is performed, ProcessedSignalError contains the error propagated from detrending. This is done by dividing the error by the normalized MultDetrendSignal. |
| gNumPixOLHi | rNumPixOLHi | integer | | Number of outlier pixels per feature with intensity > upper threshold set via the pixel outlier rejection method. The number is computed independently in each channel. These pixels are omitted from all subsequent calculations. |
| gMedianSignal | rMedianSignal | float | | Raw median signal of feature from inlier pixels in green and/or red channel |
| gPixNormIQR | rPixNormIQR | float | | The normalized Inter-quartile range of all of the inlier pixels per feature. The range is computed independently in each channel. |
| gIsSaturated | rIsSaturated | boolean | 1 = Saturated or<br>0 = Not saturated | Boolean flag indicating if a feature is saturated or not. A feature is saturated IF 50% of the pixels in a feature are above the saturation threshold. |
| gIsFeatNonUnifOL | rIsFeatNonUnifOL | boolean | g(r)IsFeatNonUnifOL = 1 indicates Feature is a non-uniformity outlier in g(r) | Boolean flag indicating if a feature is a NonUniformity Outlier or not. A feature is non-uniform if the pixel noise of feature exceeds a threshold established for a "uniform" feature. |

**Table 11** Feature results contained in the MINIMAL output text file (MINIMAL FEATURES table)

| Features (Green) | Features (Red) | Types | Options | Description |
|---|---|---|---|---|
| gIsFeatPopnOL | rIsFeatPopnOL | boolean | g(r)IsFeatPopnOL = 1 indicates Feature is a population outlier in g(r) | Boolean flag indicating if a feature is a Population Outlier or not. Probes with replicate features on a microarray are examined using population statistics.<br><br>A feature is a population outlier if its signal is less than a lower threshold or exceeds an upper threshold determined using a multiplier (1.42) times the interquartile range (i.e., IQR) of the population. |
| gIsWellAboveBG | rIsWellAboveBG | Boolean | | Boolean flag indicating if a feature is WellAbove Background or not, feature passes g(r)IsPosAndSignif and additionally the g(r)BGSubSignal is greater than 2.6*g(r)BG_SD. You can change the multiplier 2.6. |

# Other text result file annotations

The following public accession numbers may or may not show up in the Feature Results section of the output text file.

**Table 12**    Public accession numbers in the output text file

| Abbreviation | Description |
|---|---|
| dbj | DNA Database of Japan |
| emb | EMBL |
| gb | GenBank |
| gbpri | GenBank primate nucleotide accession number |
| gi | GenBank Gene Identifier |
| gp | GenPept protein identification number |
| mgi | Mouse Genome Informatics |
| pdb | Brookhaven Protein data bank |
| pir | NBRF PIR |
| prf | Protein Research Foundation |
| rafl | RIKEN full Length cDNA |
| ref | RefSeq |
| sp | SwissProt |
| tair | The Arabidopsis Information Resource |
| ug | UniGenelocuslink: LocusLink ID |
| wi | Whitehead |

# 4
# MAGE-ML (XML) File Results

This chapter provides a listing of MAGE-ML results in the form of tables. Refer to these tables when you want to know the results reported in a particular file. This chapter also contains a section on TIFF files and formats.

# How Agilent output file formats are used by databases

*Pattern files should be loaded to the database via FTP if possible to ensure that the pattern element, name attribute, is used to name the pattern.*

Data analysis programs must match up information about the layout and annotation of the microarray features with the profile result files for each microarray within their databases. Agilent provides this design information for its microarrays in a variety of file formats, including GAL and MAGE-ML. These files describe the gene probes and their number and spacing on the microarray. Profile result files contain the signal and error information for each of the hybridized gene probes on the microarray.

Both pattern files and profile result files contain information that can be formatted in several ways: tab-delimited text format or an XML format, MAGE-ML.

Agilent only supports GEML2 Pattern files and MAGE-ML profiles for use with Rosetta Resolver. The pattern name in Rosetta Resolver should match the profile pattern name embedded in the profile data so that the data can be correctly associated. To do this, use the pattern autoimport function in Rosetta Resolver or correctly specify the pattern name when manually importing the pattern. (The Agilent pattern name in most cases is "Agilent-xxxxxx" where the xxxxxx is the AMADID number of the microarray.)

For transfer of data into GeneSpring, the pattern information can be obtained from within the Feature Extraction profile tab text file or can be obtained by download from the GeneSpring Web site.

# MAGE-ML results

## Differences between MAGE-ML and text result files

The MAGE-ML result file includes most of the same parameters, statistics and results as the FULL text result file with the following differences:

- Scanner control parameters are included in the file.
- Some Feature Extraction parameter names (FE PARAMS table) have been changed to accommodate Rosetta Resolver terminology.
- MAGE result file includes all information included in the FEATURES table except for annotations, deletion control information and spot size information.
- Feature results (FEATURES table) are associated with quantitation types as defined by the Object Management Group in its Gene Expression Specification paper of February 2003 V.1. These types are listed below:
  - Measured Signal
  - Derived Signal
  - Ratio
  - Confidence Indicators—error and p-value
  - Specialized Quantitation Type (SQT) — includes all other data

## Full and Compact Output Packages

In the Properties sheet for the project you can select if you want the MAGE-ML result file to contain all the possible columns and results (Full) or a reduced set of results (Compact).

MAGE-ML files can also be compressed before they are sent via FTP. Compressed MAGE-ML files further reduces the size

of the file to decrease the transfer time. Use both Compact and Compressed MAGE-ML files for Resolver. The Compact package contains only those columns required by Resolver, GeneSpring, CGH Analytics and Chip Analytics.

In the Compact version of the MAGE-ML file, the entire FEPARAMS section is included. MAGE-ML has a rich mechanism for describing protocols and protocol parameters.

## Tables for Full Output Package

**Table 13**  Scan protocol parameters in MAGE-ML result file

| Parameter | Description |
|---|---|
| Image acquisition identifier | Barcode or identifier for microarray |
| Log information | Warnings and errors during run |
| Activity date | Time stamp for scanner run |
| Scanner information | Information such as name, make model and serial number of scanner |
| Operator | Person that runs scanner |
| ScanNumber | Number of the scan associated with the values listed in this table |
| Red.LASER_POWER_VALUE | Value of laser power in red channel |
| Green.LASER_POWER_VALUE | Value of laser power in green channel |
| Red.PMT_GAIN_VALUE | Photomultiplier gain in red channel |
| Green.PMT_GAIN_VALUE | Photomultiplier gain in green channel |
| Red.Saturation_Value | Signal value beyond which signal is saturated in the red channel |
| Green.Saturation_Value | Signal value beyond which signal is saturated in the green channel |
| MICRONS_PER_PIXEL_X | Radius of pixel in the x direction |

**Table 13**   Scan protocol parameters in MAGE-ML result file (continued)

| Parameter | Description |
|---|---|
| MICRONS_PER_PIXEL_Y | Radius of pixel in the y direction |
| GlassThickness | Thickness of microarray slide |
| Red.DarkOffsetAverage | Dark offset data per image in red channel as measured by scanner |
| Green.DarkOffsetAverage | Dark offset data per image in green channel as measured by scanner |
| PercentAutoFocusHold | Amount of movement in the autofocus because of fluctuations in the glass |
| DarkOffsetSubtracted | Resulting signal when dark offset value is subtracted |

**Table 14**   Feature Extraction protocol parameters in MAGE-ML result file Differences between FEPARAMS in text file and MAGE-ML file

| Text File FEPARAMS | MAGE-ML File FEPARAMS |
|---|---|
| Ratio_ErrorModel | Error Model |
| Ratio_AddErrorRed | Red.ADDITIVE_ERROR |
| Ratio_AddErrorGreen | Green.ADDITIVE_ERROR |
| Ratio_MultErrorRed | Red.MULTIPLICATIVE_ERROR |
| Ratio_MultErrorGreen | Green.MULTIPLICATIVE_ERROR |

**NOTE**   For 1-color, red signals and log ratios are not included in the MAGE-ML output files.

**Table 15** Feature results (Full) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| SQT[*] | X_IMAGE_POSITION<br>Y_IMAGE_POSITION | | | Found coordinates of the feature centroid |
| SQT | SpotExtentX<br>SpotExtentY | | | Diameter of the spot (X- or Y-Axis) |
| Ratio | LogRatio (base 10) | | | log(REDsignal/GREENsignal) per feature (processed signals used to calculate log ratio) |
| | | | | If SURROGATES are turned off, then: |
| | | | -4 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig > 0.0 |
| | | | 4 | if DyeNormRedSig > 0.0 & DyeNormGreenSig <= 0.0 |
| | | | 0 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig <= 0.0 |
| Error | LogRatioError | | | If SURROGATES are turned off, then: |
| | | | 1000 | if DyeNormRedSig <= 0.0 OR DyeNormGreenSig <= 0.0 |
| | | | | IF SURROGATES are turned on, then: |
| | | | | LogRatioError = error of the log ratio calculated according to the error model chosen |
| PValue | PValueLogRatio | | | Significance level of the Log Ratio computed for a feature |
| SQT | gSurrogateUsed | rSurrogateUsed | Non-zero value<br>0 | The g(r) surrogate value used<br>No surrogate value used |

**Table 15** Feature results (Full) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| SQT | gIsFound | rIsFound | 1 = IsFound<br>0 = IsNotFound | A boolean used to flag found (strong) features. The flag is applied independently in each channel.<br><br>A feature is considered found if the calculated spot centroid is within the bounds of the spot deviation limit with respect to corresponding nominal centroid. NOTE: IsFound was previously termed IsStrong. |
| Derived Signal | Green.DerivedSignal | Red.DerivedSignal | | The propagated feature signal, per channel, used for computation of log ratio |
| Error | Green.ProcessedSig Error | Red.ProcessedSig Error | | Standard error of propagated feature signal, per channel |
| SQT | gNumPixOLHi | rNumPixOLHi | | Number of outlier pixels per feature with intensity > upper threshold set via the pixel outlier rejection method. The number is computed independently in each channel. These pixels are omitted from all subsequent calculations. |
| SQT | gNumPixOLLo | rNumPixOLLo | | Number of outlier pixels per feature with intensity < lower threshold set via the pixel outlier rejection method. The number is computed independently in each channel.<br><br>NOTE: The pixel outlier method is the ONLY step that removes data in Feature Extraction. |
| SQT | gNumPix | rNumPix | | Total number of pixels used to compute feature statistics, i.e., total number of inlier pixels/per spot, same in both channels |

**Table 15**  Feature results (Full) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| Measured Signal | Green.Measured Signal | Red.Measured Signal | | Raw mean signal of feature in green (red) channel |
| SQT | gMedianSignal | rMedianSignal | | Raw median signal of feature in green (red) channel |
| SQT | gNetSignal | rNetSignal | | MeanSignal minus DarkOffset |
| Error | Green.PixSDev | Red.PixSDev | | Standard deviation of all inlier pixels per feature. This is computed independently in each channel. |
| SQT | gBGNumPix | rBGNumPix | | Total Number of pixels used to compute Local BG statistics per spot; i.e., total number of BG inlier pixels. This number is computed independently in each channel. |
| Measured Signal | Green.Background | Red.Background | | Mean local background signal (local to corresponding feature) computed per channel |
| SQT | gBGMedianSignal | rBGMedianSignal | | Median local background signal (local to corresponding feature) computed per channel |
| Error | Green.BGPixSDev | Red.BGPixSDev | | Standard deviation of all inlier pixels per Local BG of each feature, computed independently in each channel |
| SQT | gNumSatPix | rNumSatPix | | Total number of saturated pixels per feature, computed per channel |
| SQT | gIsSaturated | rIsSaturated | 1 = Saturated or 0 = Not saturated | Integer indicating if a feature is saturated or not. A feature is saturated IF 50% of the pixels in a feature are above the saturation threshold. |

**Table 15** Feature results (Full) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| SQT | gIsLowPMTScaledUp | rIsLowPMTScaledUp | 1 = Low<br>0 = High | For XDR features, this is an integer indicating if the low PMT value was used for the calculations, or the high value. |
| SQT | PixCorrelation | | | Ratio of estimated feature covariance in RedGreen space to product of feature Standard Deviation in Red Green space<br><br>The covariance of two features measures their tendency to vary together, i.e., to co-vary. In this case, it is a cumulative quantitation of the tendency of pixels belonging to a particular feature in Red and Green spaces to co-vary. |
| float | BGPixCorrelation | | | The same concept as above but in case of background |
| SQT | gIsFeatNonUnifOL | rIsFeatNonUnifOL | g(r)IsFeatNonUnifOL = 1 indicates Feature is a non-uniformity outlier in g(r) | Integer indicating if a feature is a NonUniformity Outlier or not. A feature is non-uniform if the pixel noise of feature exceeds a threshold established for a "uniform" feature. |
| SQT | gIsBGNonUnifOL | rIsBGNonUnifOL | g(r)IsBGNonUnifOL = 1 indicates Local background is a non-uniformity outlier in g(r) | The same concept as above but for background |

**Table 15** Feature results (Full) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| SQT | gIsFeatPopnOL | rIsFeatPopnOL | g(r)IsFeatPopnOL = 1 indicates Feature is a population outlier in g(r) | Boolean flag indicating if a feature is a Population Outlier or not. Probes with replicate features on a microarray are examined using population statistics.<br><br>A feature is a population outlier if its signal is less than a lower threshold or exceeds an upper threshold determined using a multiplier (1.42) times the interquartile range (i.e., IQR) of the population. |
| SQT | gIsBGPopnOL | rIsBGPopnOL | g(r)IsBGPopnOL = 1 indicates local background is a population outlier in g(r) | The same concept as above but for background |
| SQT | IsManualFlag | | | |
| SQT | gBGSubSignal | rBGSubSignal | gBGSubSignal = gMeanSignal - gBGUsed | Background-subtracted signal To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| Error | gBGSubSigError | rBGSubSigError | | Propagated standard error as computed on net g(r) background-subtracted signal |
| SQT | BGSubSigCorrelation | | | Ratio of estimated background-subtracted feature signal covariance in RG space to product of background-subtracted feature Standard Deviation in RG space |

**Table 15** Feature results (Full) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| SQT | gIsPosAndSignif | rIsPosAndSignif | g(r)isPosAndSignif = 1 indicates Feature is positive and significant above background | Boolean flag, established via a 2-sided t-test, indicates if the mean signal of a feature is greater than the corresponding background (selected by user) and if this difference is significant. To display variables used in the t-test, see Table 20 on page 190. |
| SQT | gPValFeatEqBG | rPValFeatEqBG | | P-value from t-test of significance between g(r)Mean signal and g(r) background |
| SQT | gIsWellAboveBG | rIsWellAboveBG | | Boolean flag indicating if a feature is WellAbove Background or not Feature passes g(r)IsPosAndSignif and additionally the g(r)BGSubSignal is greater than 2.6*g(r)BGSDUsed. |
| Boolean | gSpatialDetrendIsIn FilteredSet | rSpatialDetrendIsIn FilteredSet | | Set to true for a given feature if it is part of the filtered set used to detrend the background. This feature is considered part of the locally weighted lowest x% of features as defined by the *DetrendLowPassPercentage*. |
| float | gSpatialDetrend SurfaceValue | rSpatialDetrend SurfaceValue | | Value of the smoothed surface calculated by the Spatial detrend algorithm |
| SQT | IsUsedBGAdjust | | 1 = Feature used 0 = Feature not used | A boolean used to flag features used for computation of global BG offset |
| SQT | gBGUsed | rBGUsed | gBGSubSignal = gMeanSignal - gBGUsed | Background used to subtract from the MeanSignal; variable also used in t-test. To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |

\* SQT — Specialized Quantitation Type

# Table for Compact Output Package

This table contains only those columns required by Resolver, GeneSpring, CGH Analytics and Chip Analytics.

In the Compact version of the MAGE-ML file, the entire FEPARAMS section is included. MAGE-ML has a rich mechanism for describing protocols and protocol parameters.

**Table 16**   Feature results (Compact) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| Ratio | LogRatio (base 10) | | | log(REDsignal/GREENsignal) per feature (processed signals used to calculate log ratio) |
| | | | | If SURROGATES are turned off, then: |
| | | | -4 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig > 0.0 |
| | | | 4 | if DyeNormRedSig > 0.0 & DyeNormGreenSig <= 0.0 |
| | | | 0 | if DyeNormRedSig <= 0.0 & DyeNormGreenSig <= 0.0 |
| SQT[*] | X_IMAGE_POSITION Y_IMAGE_POSITION | | float | Found coordinates of the feature centroid in microns |
| Error | LogRatioError | | | If SURROGATES are turned off, then: |
| | | | 1000 | if DyeNormRedSig <= 0.0 OR DyeNormGreenSig <= 0.0 |
| | | | | IF SURROGATES are turned on, then: |
| | | | | LogRatioError = error of the log ratio calculated according to the error model chosen |
| PValue | PValueLogRatio | | | Significance level of the Log Ratio computed for a feature |

**Table 16** Feature results (Compact) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| Derived Signal | Green.DerivedSignal | Red.DerivedSignal | | The propagated feature signal, per channel, used for computation of log ratio |
| Error | Green.ProcessedSig Error | Red.ProcessedSig Error | | Standard error of propagated feature signal, per channel |
| Measure d Signal | Green.Measured Signal | Red.Measured Signal | | Raw mean signal of feature in green (red) channel |
| SQT | gMedianSignal | rMedianSignal | | Raw median signal of feature in green (red) channel |
| SQT | gBGMedianSignal | rBGMedianSignal | | Median local background signal (local to corresponding feature) computed per channel |
| Error | Green.BGPixSDev | Red.BGPixSDev | | Standard deviation of all inlier pixels per Local BG of each feature, computed independently in each channel |
| SQT | gIsSaturated | rIsSaturated | 1 = Saturated or 0 = Not saturated | Integer indicating if a feature is saturated or not. A feature is saturated IF 50% of the pixels in a feature are above the saturation threshold. |
| SQT | gIsLowPMTScaledUp | rIsLowPMTScaledUp | 1 = Low 0 = High | For XDR features, this is an integer indicating if the low PMT value was used for the calculations, or the high value. |
| SQT | gIsFeatNonUnifOL | rIsFeatNonUnifOL | g(r)IsFeatNonUnifOL = 1 indicates Feature is a non-uniformity outlier in g(r) | Integer indicating if a feature is a NonUniformity Outlier or not. A feature is non-uniform if the pixel noise of feature exceeds a threshold established for a "uniform" feature. |

**Table 16** Feature results (Compact) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| SQT | gIsBGNonUnifOL | rIsBGNonUnifOL | g(r)IsBGNonUnifOL = 1 indicates Local background is a non-uniformity outlier in g(r) | The same concept as above but for background |
| SQT | gIsFeatPopnOL | rIsFeatPopnOL | g(r)IsFeatPopnOL = 1 indicates Feature is a population outlier in g(r) | Boolean flag indicating if a feature is a Population Outlier or not. Probes with replicate features on a microarray are examined using population statistics.<br><br>A feature is a population outlier if its signal is less than a lower threshold or exceeds an upper threshold determined using a multiplier (1.42) times the interquartile range (i.e., IQR) of the population. |
| SQT | gIsBGPopnOL | rIsBGPopnOL | g(r)IsBGPopnOL = 1 indicates local background is a population outlier in g(r) | The same concept as above but for background |
| SQT | gBGSubSignal | rBGSubSignal | gBGSubSignal = gMeanSignal - gBGUsed | Background-subtracted signal To display the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| SQT | IsManualFlag | | | Boolean flag that describes if the feature centroid was manually adjusted. |

**Table 16** Feature results (Compact) contained in the MAGE-ML (FEATURES table)

| Quant Type | Features (Green) | Features (Red) | Options | Description |
|---|---|---|---|---|
| SQT | gIsPosAndSignif | rIsPosAndSignif | g(r)isPosAndSignif = 1 indicates Feature is positive and significant above background | Boolean flag, established via a 2-sided t-test, indicates if the mean signal of a feature is greater than the corresponding background (selected by user) and if this difference is significant. To display variables used in the t-test, see Table 20 on page 190. |
| SQT | gIsWellAboveBG | rIsWellAboveBG | | Boolean flag indicating if a feature is WellAbove Background or not Feature passes g(r)IsPosAndSignif and additionally the g(r)BGSubSignal is greater than 2.6*g(r)BGSDUsed. |

\* SQT — Specialized Quantitation Type

# Helpful hints for transferring Agilent output files

## XML output

There are several situations you should be aware of as you use MAGE-ML (XML) output with gene expression data analysis software from Rosetta BioSoftware (Rosetta Resolver software):

### If there is no barcode

If there is no barcode in the original .tif file for whatever reason, there will be no barcode information in the MAGE-ML output (warning message in Project Run summary). For the data to load into Rosetta Resolver, it must have a barcode associated with it. You can add barcode information in the Scan Image Properties dialog box. See the *Agilent Feature Extraction for CytoGenomics User Guide*.

### Access control list (ACL)

Rosetta Resolver knows about the access control list (ACL) assigned to the scan and can easily recognize and load any MAGE-ML file. The owner of the data sets the chip and hybe access controls in Rosetta Resolver before importing the profile (scan) data. For autoimport, the profile is normally placed in the **MAGE** directory.

### XML Control Type output

If a feature is used in dye normalization, its Control_Type is normalization, even though it can also be a positive or negative control. If a feature is not used in normalization, it is either positive, negative, deletion, mismatch, or false.

**Table 17** Control Type Definitions

| Name | XML |
| --- | --- |
| Probe | false |
| Positive Control | pos or positive |
| Negative Control | neg or negative |
| Not Probe* | notprobe |

*Not Probe—These features are feature extracted, but they are not used by Feature Extraction as input to any calculations; these features are not used during outlier analysis or for the dye normalization calculation. However, dye normalization values and ratios are calculated, and the results appear in the text and XML output files, and the feature extraction visual results file. An exception is that Not Probe's background is used in the calculation of the local background with the radius method.

### Conversion of feature flag information

Failed (MAGE-ML) produce the following settings:

- Bit 8 (green) and 12 (red) are set if the feature is saturated in both channels.
- Bit 18 is set if the feature, or its deletion control, is a non-uniformity outlier in either color, or if the feature is a population outlier in either color and the Report Population Outliers as Failed in MAGE-ML file option is set to True.
- Bit 23 is set if the probe is low specificity, e.g., when the deletion control is greater than or equal to the feature.

# TIFF Results

You can transfer the original TIFF file or a JPEG file to Rosetta Resolver or a third-party program. The shape file, .shp, created during Feature Extraction cannot be displayed by any program other than Agilent Feature Extraction software.

## TIFF file format options

*See the Agilent Feature Extraction for CytoGenomics User Guide for more information on the File Info dialog box.*

Feature Extraction supports the TIFF file format. All file information for each file is listed in the File Info dialog box. The TIFF file is compliant with Adobe version 6.0 file format. The complete specification is available from the following URL:
http://partners.adobe.com/asn/developer/PDFS/TN/TIFF6.pdf.

There are two sets of custom TIFF tags in the Agilent file format.

### Genetic Analysis Technology Consortium (GATC) TIFF Tags

Agilent Technologies is not a member of GATC or otherwise connected to this organization, and makes no internal use of these tags. They are included for the convenience of customers who use software that requires them.

**Custom TIFF Tags**   Agilent Technologies uses its own custom TIFF tags for storing additional file information.

**TIFF Tag 37701**   This tag points to a data structure. This data structure is not public, but information stored in the data structure is available to customers in the MATLAB file format.

**TIFF Tag 37702**   This tag points to a string containing the file description. The usual TIFF description tags (tag 270) are used to hold the color name, "red" or "green," for each image. This allows programs that interpret only "standard" TIFF tags to determine image colors. The Page Name tag (tag 285) also contains the color names.

# 5
# How Algorithms Calculate Results

This chapter shows you how each Feature Extraction algorithm uses its parameters to calculate results that are passed on to the next algorithm and finally on to third-party data analysis programs.

**Agilent Technologies**

# Overview of Feature Extraction algorithms

Protocol step algorithms operate similarly during the Feature Extraction process for 2-color gene expression, CGH, ChIP, and non-Agilent microarrays. That is, the algorithms and parameter fields are similar, but the parameter values are different depending on the protocol.

The Feature Extraction process for 1-color gene expression microarrays includes only seven protocol steps, and for miRNA analysis the process includes those seven steps plus a MicroRNA Analysis step.

The examples used below are primarily for 2-color microarrays. Any differences in algorithms and functions for other microarray experiments are also explained.

## Algorithms and functions they perform

### Place Grid

This algorithm finds the grid to define the nominal positions of the spots on the microarray.

*For more information on the algorithms for XDR extraction, see "XDR Extraction Process" on page 170.*

**eXtended Dynamic Range (XDR) extraction**    For an XDR extraction, the grid placement is done using the high intensity scan (i.e., higher PMT voltage). The grid found using the high intensity scan is used as the starting point for the remaining extraction of both the high and low intensity images.

With version 10.x and higher of the software, you no longer have to perform XDR dual scans or extractions to capture the full dynamic range of the data. You can get the same dynamic range by working with the 20-bit TIFF Dynamic Range option. This option is meant to be a replacement for the XDR option. You capture the full dynamic range with better accuracy.

Choosing the XDR option may still be useful if you want to compare XDR data from the G2565BA Scanner with XDR data from the G2565CA Scanner.

### Optimize Grid Fit

This algorithm improves the grid fit on the entire microarray. Leveraging from the Spot Finder algorithm, this protocol step examines the spots in the four corners of the microarray and iteratively adjusting the grid for a better fit.

If the grid has been optimized by this protocol step, the STATS table shows the stat GridHasBeenOptimized with boolean of 1; or a boolean of 0 if the grid has not been optimized.

### Find Spots

This algorithm locates the exact size and centroid of each spot on the scanned microarray. Once the spot centroids have been located, the CookieCutter algorithm or WholeSpot algorithm defines the feature for each spot. The software then defines the local background for each spot based on the radius of a circle drawn around the spot.

Next, the pixel outlier algorithm identifies outlier pixels in the feature and in the local background for each spot. These pixels are then omitted from further calculations. This is the only point where data is omitted. Subsequent outlier analyses flag data, but do not remove the data.

Inlier pixels within the cookie area represent a feature while the inlier pixels within the annulus around the feature, after excluding the exclusion zone, represent the local background. The Feature Extraction program calculates the following

values from these inlier pixels: mean, median, standard deviation, normalized IQR, and number of inlier pixels.

**XDR extraction**   This is the only step that is run twice on an XDR extraction. The spot placement and spot measurements are found separately for the high and low intensity scans. Then the XDR algorithm decides on a feature by feature basis which scan the data should come from (more on this below). For features that are very bright in the high intensity scan, the XDR algorithm uses the data from the low intensity scan. This choice is made independently for each color channel.

For each feature that uses data from the low intensity scan, the following columns get replaced (determined separately for red and green channels): NumPixOLHi, NumPixOLLo, NumPix, MeanSignal, MedianSignal, PixSDev, PixNormIQR, NumSatPix, IsSaturated, NetSignal.

These columns include the raw data from the spotfinding and measurement steps (signal levels, pixel noise levels, number of pixels, if the pixels and feature are saturated). Once the substitutions have been made to some features in each color channel, the extraction proceeds as if there were only a single combined set of features.

### Flag Outliers

Next, the Flag Outliers algorithm flags anomalous features and local backgrounds as non-uniformity outliers and/or population outliers. Population outlier flagging is based on population statistics of replicate features on the microarray.

Which of two statistical tests is used to identify population outliers depends on the number of replicate features on the microarray.

Non-uniformity outlier flagging is based on statistical deviation from the expected noise in the Agilent microarray-based system (scanner, labeling/hybridization protocols, and microarrays). The algorithm automatically calculates the B (linear) and C (constant) terms of the

polynomial fit for the expected noise for any type of microarray experiment.

### Compute Bkgd, Bias and Error

This algorithm applies background subtraction to each feature to yield the background-subtracted intensity. You can also apply a "spatial detrend" algorithm to estimate and remove noise due to a systematic gradient on the microarray.

Another algorithm can correct for any underestimation or overestimation of the background in both the red and green channels of low-intensity signals by applying a global background adjustment value to the background-subtracted signals.

Before using the algorithm for estimating the error, the system uses an algorithm to calculate robust negative control statistics for both CGH and miRNA data.

CGH microarrays have a variety of sequences that are used as negative controls. Occasionally, "hot" features are not flagged as population outliers. In addition, "hot" sequences may exist; that is, all features of that sequence have higher signals than features in other negative control sequences. These problems can inflate NegC SD, which is used in the calculation of AdditiveError for the CGH error model.

To provide an estimate of the error in the background-subtracted signal calculation, the error model is now calculated after background subtraction. The 1-color error model has been changed to exactly mimic the 2-color error model.

To determine if the feature intensity is significant compared to the background intensity, two kinds of tests are available: t-test and WellAboveBG test. Both of these tests depend upon an estimation of background error.

The default protocol for older Agilent protocols still uses pixel statistics of local background regions to estimate background error in the 2-sided t-test. Newer Agilent protocols use an improved estimation of background error:

the additive error, calculated from the Agilent error model. You can choose between these two background error estimations in the protocol parameter field, "Significance (for IsPosAndSignif and IsWellAboveBG)".

The WellAboveSDMulti confidence test is used to determine if the feature background-subtracted signal is well above its background error.

Surrogates are calculated here and depend on the significance model used. Given the standard t-test, the surrogates are calculated exactly as before. Given the new significance test based upon additive error, the surrogate value is determined by the additive error and the p-value.

The program can also use a multiplicative detrend algorithm, if selected or the default in the protocol, to provide a surface fit to account for the dome effect that can happen when microarrays are processed.

Placing the error model calculation step before the significance calculation permits the result of the error model calculation to be used for the significance calculation, surrogate calculation and multiplicative detrending steps.

### Correct Dye Biases

Since dye bias between the red and green channels is a common phenomenon in a dual-color microarray platform, this algorithm adjusts for the bias by multiplying the background-subtracted signals with the appropriate dye normalization factors. Both linear and non-linear (locally weighted) normalization methods are available.

Surrogates are applied after the dye norm fit and before the dye normalization takes place. This ensures that only real data contribute to the fit and also surrogate data is correctly dye-normalized for both the Linear and Lowess options.

Because 1-color experiments use only the green channel, they do not use this protocol step. Surrogates exist and can be used for 1-color.

### Compute Ratios

This algorithm determines if a feature is differentially expressed by calculating the log ratio of the red over green processed signals. The processed signal is the dye-normalized signal.

Because 1-color experiments use only the green channel, they do not use this protocol step.

### Calculate Metrics

These algorithms calculate all the QC metrics for the analysis. One of the primary algorithms in this step is the gridding test, whose parameter values are hidden in the protocol. This algorithm yields grid warnings on the Summary Reports and the "Evaluate Grid" warning in the QC Report. Agilent has added many more tests to assess if gridding has been successful or not.

Protocols for Agilent arrays also have associated QC metric sets. These metrics are calculated at this step.

### Generate Results

This part of the process generates the output result files using the parameter values specified in the protocol step and the selections made in the Project Properties window. This step is not discussed in this chapter.

# Algorithms and results they produce

The table below summarizes the results for each algorithm (protocol step). These result names are used in the equations for the calculations for each algorithm.

**Table 18**  Algorithms (Protocol Steps) and the results they produce

| Protocol Step | Results | Result Definition |
|---|---|---|
| Find Spots | MeanSignal | Average raw signal of feature calculated from the intensities of all inlier pixels that represent the feature (after outlier pixel rejection). The number of inlier pixels is shown in the column NumPix. |
| Find Spots | MedianSignal | Median raw signal of feature calculated from the intensities of all inlier pixels that represent the feature (after outlier pixel rejection). The number of inlier pixels is shown in the column NumPix. |
| Find Spots | BGMeanSignal | Average raw signal of the local background calculated from intensities of all inlier pixels that represent the local background of the feature (after outlier pixel rejection). The number of inlier pixels is shown in the column BGNumPix. |
| Find Spots | BGMedianSignal | Median raw signal of the local background calculated from intensities of all inlier pixels that represent the local background of the feature (after outlier pixel rejection). The number of inlier pixels is shown in the column BGNumPix. |
| Find Spots | NetSignal | MeanSignal minus Dark Offset |
| Find Spots | IsSaturated | A Boolean flag of 1 indicates that the feature is saturated; at least 50% of the inlier pixels in the feature have intensities above the saturation threshold. One can determine the saturation level of a feature by dividing the NumSatPix by the NumPix. |
| Flag Outliers | IsFeatureNonUnifOL | A Boolean flag of 1 indicates that the feature is a non-uniformity outlier; the measured feature pixel variance is greater than the expected feature pixel variance plus the confidence interval. |
| Flag Outliers | IsFeatPopOL | A Boolean flag of 1 indicates that the feature is a population outlier. This means that the feature MeanSignal is greater than the upper rejection boundary or less than the lower rejection boundary, both of which are determined by multiplying a factor (1.42) by the interquartile range of the population, made up of intra-array feature replicates. (See "Step 6. Reject outliers" on page 181.) |

**Table 18** Algorithms (Protocol Steps) and the results they produce (continued)

| Protocol Step | Results | Result Definition |
|---|---|---|
| Compute Bkgd, Bias and Error | BGAdjust | An adjustment value added to the initial background-subtracted signal to correct for underestimation or overestimation of the background. This value can be positive or negative. Note the BGAdjust values are reported per channel in the STATS table of Feature Extraction text file. |
| Compute Bkgd, Bias and Error | BGused | Final background signal used to subtract the background from the feature mean signal. To view the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| Compute Bkgd, Bias and Error | BGSubSignal | Feature signal after subtraction of the background corrections. To view the values used to calculate this variable using different background signals and settings of spatial detrend and global background adjust, see Table 20 on page 190. |
| Compute Bkgd, Bias and Error | IsPosAndSignif | If significance is based on pixel statistics, a Boolean flag of 1 indicates that the feature MeanSignal is greater than and significant compared to the background signal (i.e BGUsed). |
| | | If significance is based on the Additive Error of the Error Model, a Boolean flag of 1 means that the feature MeanSignal is greater than and significant compared to the Additive Error, |
| Compute Bkgd, Bias and Error | IsWellAboveBG | A Boolean flag of 1 indicates that the feature BGSubSignal is well above background and passes the IsPosAndSignif test. |
| Compute Bkgd, Bias and Error | SpatialDetrendIsIn FilteredSet | Set to true for a given feature if it is part of the filtered set used to detrend the background. The feature may be in the set of locally weighted lowest x% of features as defined by the *DetrendLowPassPercentage,* may be a negative control feature or may be part of the set of features that are in the negative control range. The feature set is defined by the detrend method selected. |
| Compute Bkgd, Bias and Error | SpatialDetrend SurfaceValue | Value of the smoothed surface, at that feature, calculated by the Spatial detrend algorithm |

**Table 18** Algorithms (Protocol Steps) and the results they produce (continued)

| Protocol Step | Results | Result Definition |
|---|---|---|
| Compute Bkgd, Bias and Error | MultDetrendSignal | A surface is fitted through the log of the background-subtracted signal to look for multiplicative gradients. A normalized version of that surface interpolated at each point of the microarray is stored in MultDetrendSignal. The surface is normalized by dividing each point by the overall average of the surface. That average is stored in MultDetrendSurfaceAverage as a statistic.

If the protocol uses the option to fit to only replicate features, the surface is normalized for the fit. The MultDetrend SurfaceAverage is smaller in this case, a number around 1. |
| Compute Bkgd, Bias and Error | SurrogateUsed | A non-zero surrogate value indicates that the MeanSignal is less than or not significant versus the background or the BGSubSignal is less than the Error, where the Error is the Additive Error for all default Agilent Protocols. |
| Correct Dye Biases | DyeNormSignal | A dye-normalized signal calculated by multiplying the BGSubSignal with the appropriate DyeNormFactor. |
| Correct Dye Biases | LinearDyeNormFactor (Table 3 on page 71) | A global constant to normalize the dye bias from all feature background-subtracted signals. LinearDyeNormFactor is calculated such that geometric mean intensity of the selected normalization features equals 1000. |
| Compute Ratios | ProcessedSignal | The signal left after all the Feature Extraction processing steps have been completed. In the case of 1-color, ProcessedSignal contains the Multiplicatively Detrended BackgroundSubtracted Signal if the detrending is selected and helps. If the detrending does not help, this column will contain the BackgroundSubtractedSignal. |
| Compute Ratios | ProcessedSigError | The universal or propagated error left after all the processing steps of the Feature Extraction process have been completed. In the case of one color, If multiplicative detrending is performed, ProcessedSignalError contains the error propagated from detrending. This is done by dividing the error by the normalized MultDetrendSignal. |
| Compute Ratios | LogRatio | Log of the ratio of rProcessedSignal over gProcessedSignal. The log ratio indicates the level of gene expression in cyanine 5-labeled sample relative to cyanine 3-labeled sample. |

**Table 18** Algorithms (Protocol Steps) and the results they produce (continued)

| Protocol Step | Results | Result Definition |
|---|---|---|
| Compute Ratios | pValueLogRatio | P-value indicates the level of significance in the differential expression of a gene as measured through the log ratio. |
| MicroRNA Analysis | gTotalGeneSignal | This signal is the sum of the total probe signals in the green channel per gene. |
| MicroRNA Analysis | gTotalGeneError | This error is the square root of the sum of the squares of the TotalProbeError. |

# XDR Extraction Process

## What is XDR scanning?

The Agilent scanner can cover a dynamic intensity range greatly in excess of the range covered by a single scan. Furthermore, Agilent microarray features can produce signals that span a broader range of intensity than a single scan can cover. Therefore, you can use eXtended Dynamic Range (XDR) to cover the full dynamic intensity range of your microarray features and hence see the most useful biology.

To do this you set the scanner to scan twice, once at a high PMT setting (the high intensity scan) followed immediately by a low PMT setting (the low intensity scan). This functionality is enabled using Agilent Scan Control Software version 7.0. The two scans are labeled in their tiff headers as paired scans of the same microarray.

## XDR Feature Extraction process

The Feature Extraction program (v9.1 and later) uses this information to know to extract the low and high PMT images as a pair. In this XDR extraction type, the Feature Extraction program processes the two scans together and produces a single set of outputs that contain data from both scans.

Some of the features contain data from the high intensity scan and some from the low intensity scan. You can determine this by viewing the column, r,gIsLowPMTScaledUp, for each color channel. For signals that are very bright (or saturated) in the high intensity scan (e.g., a scan at 100% PMT gain), the XDR algorithm substitutes the data from the low intensity scan (e.g., 10% PMT gain) after scaling the intensity appropriately.

To extract these arrays, the Feature Extraction program uses a somewhat different flow of the image processing and data analysis algorithms.

The Feature Extraction program places the grid on the high intensity scan only, then finds spots using this grid on each of the two scans.

The XDR algorithm decides which features should use the low intensity scan data, scales these signals appropriately and does a replacement for each feature and color channel where appropriate. Then Feature Extraction proceeds with the rest of the data analysis (outlier detection, background correction, dye normalization, etc.) exactly as it would for a single non-XDR scan.

Upon completion, the Feature Extraction program generates results as if they were from a single measurement of the microarray. The QC report and the stats table indicate that the Feature Extraction program extracted an XDR image pair by stating the new saturation value. This is the saturation value of the low intensity scan after suitable scaling. For instance, if the high intensity scan is at 100% and the low intensity scan is at 10%, the new saturation values will be around 650,000 (about 10x greater than a normal 100% PMT gain scan). This lets you use data in your calculations covering a much greater dynamic range.

# How the XDR algorithm works

How does the XDR algorithm decide how to combine and scale the data from the high intensity and low intensity scans? The general theory is that the high intensity gives the best results for the low end of the signal range and the low intensity scan gives better data for bright features (less affected by saturation). The Feature Extraction program uses a signal level of 20,000 as the cut-off between the two scans. If the NetSignal of the high intensity scan is greater than 20,000 counts, then the data from the low intensity scan is used.

The low intensity scan is scanned with a lower PMT gain than the high intensity scan (say 10% versus 100%). So to combine the data, the signals from the low intensity scan must be increased to match those from the high intensity scans.

To determine the factor by which the low-intensity signal should be scaled, the algorithm uses features that have signals in an overlap range where both the high and low intensity scans provide very stable data. This range is Net Signals in the high intensity scan greater than 300 counts and less than 20,000 counts.

Using data in this range, the Feature Extraction program generates a linear fit (with a slope and an intercept) that transforms the low-intensity mean signals into the same range as high intensity scans. The final scaled signal for the XDR extraction is MeanSignal ([low-intensity scan * slope] + intercept).

The linear fit constants determined in this step are included in the stats table.

For signals over 20,000 counts in the high intensity scan, therefore, the low intensity scan signals can extend to nearly 1.2 million counts.

If the low intensity scan has a spot centroid too far from the high intensity centroid (greater than 2 pixels), the algorithm does not make a substitution.

# Troubleshooting the XDR extraction

The XDR algorithm provides warnings in the project summary report to indicate an issue with the XDR extraction process.

- No XDR signal substitution for color red/green.

    This message appears if there are no features for which the low intensity data are substituted. This could occur on a dim array

Computation of the XDR fit for red/green is based on only X pairs of (high PMT, low PMT) matching values.

    This message appears if very few features had data in the overlap range for the fit. The user should check the data in this case to confirm that the XDR combination is satisfactory.

- Computation of the XDR fit for red/green results in a large intercept.

    This message appears if the linear fit between the low and high intensity scans has a very large intercept.

    This can be indicative of a poor linear fit. The user should check the data in this case to confirm that the XDR combination is satisfactory.

- Computed XDR ratio for red/green is X vs. expected Y from PMT settings. Check scanner calibration.

    This message appears if the ratio of the high/low intensity scans is different from what is expected from the scanner. For instance, an XDR scan set with 100% and 10% for PMT gain settings should yield a ratio close to 10.

    If this ratio is different than expected, the Feature Extraction program may or may not have performed correctly. But you should check the data in this case to confirm that the XDR combination is satisfactory.

    This message is more likely to appear as the low intensity PMT gain setting gets closer 1%. This is because the percentage error in the PMT gain setting increases as the setting moves away from 100%.

# How each algorithm calculates a result

## Place Grid

### Step 1. Place a grid to find the nominal spot positions

After the Feature Extraction program automatically determines the format of the grid, it initiates the next steps.

The algorithm reduces the two-dimensional image data of the microarray to two one-dimensional data sets that are further processed to determine the layout of the grid on the microarray.

Projection of the two-dimensional microarray is performed to produce two one-dimensional data sets (projected signals). From the one-dimensional data sets, peaks of the projected signals are filtered to determine which peaks to retain for further processing, based on predetermined peak height and peak width thresholds.

Nominal spacing between the features may be estimated based on a statistical determination of a most frequent distance between centers of retained peaks that are adjacent to one another. Coordinates for the features on the microarray, relative to the X and Y axes, are generated based on the selected peaks and peak spacing. The grid is then adjusted for rotation and skew

.

**NOTE**    In Feature Extraction for CytoGenomics 5.0, an enhanced gridding algorithm was released and used in the preloaded protocols. The enhancements include a new iterative method for determining grid position, rotation, and skew, and several "fine" grid tuning methods that improve the calculation of rotation and skew. Enhanced gridding also uses both the foreground and background of the corner stencil patterns to improve identification of grid corners.

The background peak shift flag helps to improve the gridding. Ideally, all background pixels should have a gray value of zero. In practice these values are nonzero.

When this flag is set to true, the algorithm determines the background pixels' pixel value from the histogram of the image. All pixels having a non-zero value (background +/- window) are set to zero, thus reducing the contribution of background pixels in the two one-dimensional projected signals. This shift in the peak of the background signal leads to better determination of peaks.

The following figures illustrate the result of applying Background Peak Shifting. Figure 37 is a histogram of a typical 30 micron feature array before Background Peak Shifting. Figure 38 depicts the same array after applying Background Peak Shifting. Note that this operation is done internally in the grid placement algorithm. The actual image data remains unchanged. Some variations in the results are expected with and without use of this flag as the grid positions obtained differ.
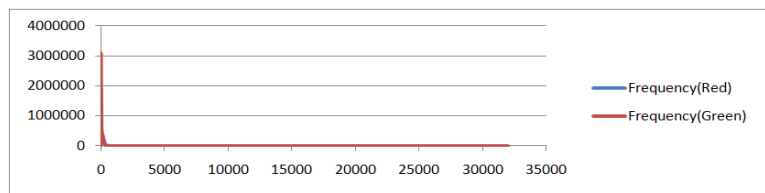


**Figure 37**   Histogram of a 30 micron feature array image. The X-axis corresponds to the pixel value and the Y-axis to the frequency of occurrence.
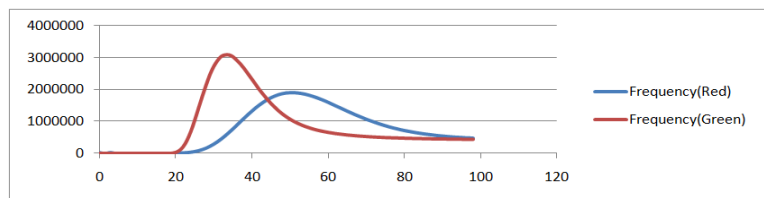


**Figure 38**   Zoomed in section of Figure 37. The background peaks are at 32 for the red channel and 50 for the green channel.
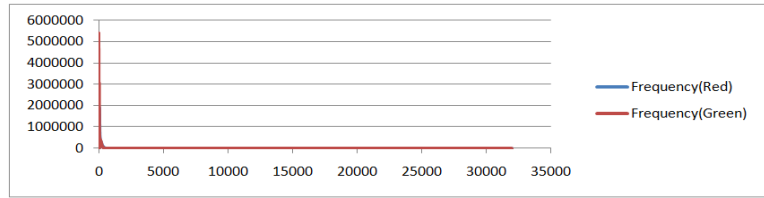
**Figure 39**    Histogram of a 30 micron feature array image after Background Peak Shifting.
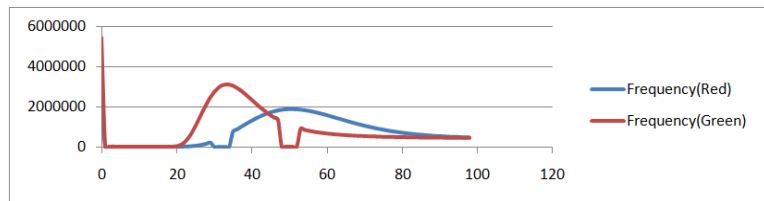


**Figure 40**    Zoomed in section of Figure 39. Note the peaks at pixel value=0. Also note the dips in the frequency of values near the pixel value of 32 for the red channel and 50 for the green channel.

When the Use central part of pack for slope and skew calculation flag is set to True, the gridding algorithm is modified to use central region of the pack to obtain slope, skew and origin of each pack, instead of using the edges of packs. This enables the algorithm to correctly place the grid for arrays that have edges populated with dim spots.

When the Use the correlation method to obtain origin X of subgrids is set to False, results obtained from the projection data analysis are used to estimate the origin. Selecting this option will use the same calculations used in Feature Extraction version 10.7/10.9 or earlier. When the flag is set to True, the software performs one extra step of correlation following the projection data analysis to get the origin. This option is of use particularly in cases where pack edges have dim spots and are failing to grid.

# Optimize Grid Fit

### Step 2. Iteratively adjust grid by examining the corner spots

This algorithm improves the grid fit by leveraging from the Spot Finder algorithm. Looking only at the specified square area of features at each corner of the microarray, it performs the iteratively adjust corners method up to the maximum number of iterations specified in the protocol. It adjusts the grid only if the following criteria are met.

- The absolute average difference between the grid position and the spot position is within the specified Adjustment Threshold.

- The number of features considered found by the spot finder algorithm is within the specified Found Spot Threshold.

# Find Spots

### Step 3. Locate the spot centroids

The calculation is based on an iterative Bayesian-probability-based pixel classification. A binary feature mask is created that classifies the pixels in a region of interest around each grid position into feature pixels or background pixels. The approximate radius of each feature mask is considered as the corresponding spot radius and the center of mass of the feature mask is considered as the actual spot centroid.

In the visual results view (.shp file), all spots that are found are shown using a blue "X" on the spot and marked as "Found". For all spots, the blue cross (+) shows the location of the grid. If the centroid cannot be found because the spot is too weak, or the distance between + and X centroids exceeds the range specified by the Spot Deviation Limit, this spot is labeled "Not Found".

### Step 4. Define features

See the *Agilent Feature Extraction for CytoGenomics User Guide* for how the Feature Extraction program defines features either with the CookieCutter method or the WholeSpot method.

### Step 5. Estimate the radius for the local background

The radius is the distance from the center of the cookie or whole spot to the edge of the outermost region, as shown in Figure 41. The default radius is the value specified in the protocol. You can also enter a minimum radius whose value is less than the default radius, or you can enter a larger radius to capture more pixels in the background. You can use the radius method for estimating global backgrounds as well.

The figures in this step represent the local background for the CookieCutter method for defining features. The radius for the local background is estimated in the same way for the WholeSpot method.
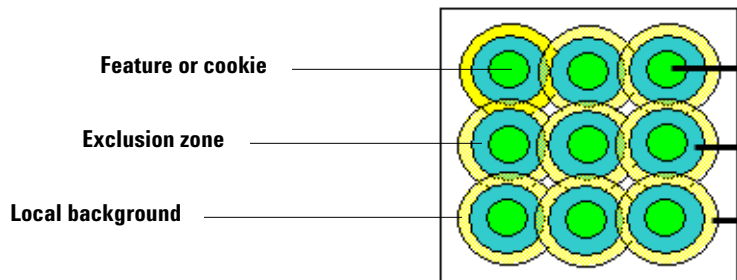


**Figure 41**    Local background in relation to other zones for CookieCutter method

**Default radius**    The default radius is the radius of the local background for one feature. This radius is known as the SELF radius and its value is the default value that you see in the Find and Measure Spots protocol step if autoestimation is turned off.

*Although the radius can map a circle that appears to overlap other features, the Feature Extraction program does not use these pixels to calculate the local background signal.*
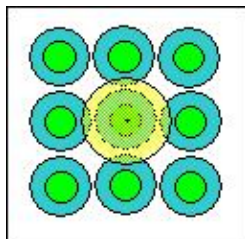


**Figure 42**     Example of a SELF radius

The value of the default radius (in microns) depends on the scan resolution and interspot spacing found in the TIFF and grid template or file, shown in equation [1]:

*Default Local Radius = SELF = (0.6* x *Scan_resolution* x *Max (Interspotspacing_x, Interspotspacing_y))* **[1]**

*For the WholeSpot method, if extraction stops at this step, you may need to enter a larger radius than the protocol default radius.*

The software autoestimates the Default Local Radius if specified in the protocol. Otherwise, you can enter this radius in the Feature Extraction Protocol Editor.

**Minimum radius**    The minimum radius that you can enter is the FLOOR (Default Radius), where FLOOR rounds the calculated value of the default radius down to the next lower integer, e.g., FLOOR (87.6) = 87.

**Maximum radius**    The software lets you enter a maximum radius for the local background no greater than the distance from the center of the innermost feature to the edge of a circle that approximately surrounds the fourth closest set of nearest neighbors, or *n*=4, as shown in Equation 2. The set of eight nearest neighbors closest to the feature of interest is defined as *n*=1, as shown in Equation 3.
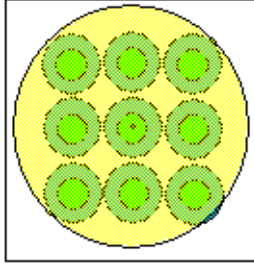
**Figure 43**    Example of the radius for the first closest set of nearest neighbors, or *n*=1 (eight nearest neighbors)

The value of the maximum radius also depends on the scan resolution and interspot spacing in the TIFF and grid template or file, shown in the equation below.

$$Max\ radius = \text{CEILING}\ [(Scan\_resolution \times 4.7)\ \sqrt{(Interspotspacing\_x)^2 + (Interspotspacing\_y)^2}] \qquad \textbf{[2]}$$

where CEILING rounds the calculated value up to the next higher integer, e.g., CEILING [3.2] = 4.

**Any radius**    The value of any radius between the minimum and maximum that circumscribes a circle surrounding the *n*th closest set of nearest neighbors from the central spot can be approximated as:

$$Radius\_n = Scan\_resolution \times n.6\ \sqrt{[(Interspotspacing\_x)^2 + (Interspotspacing\_y)^2]} \qquad \textbf{[3]}$$

where *n*=1,2,3 or 4. Figure 44 shows the set of nearest neighbors where $n = 2$.

24 nearest
neighbors
(n = 2)

2
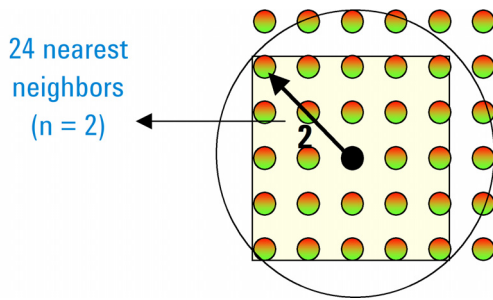
**Figure 44**   Example of the radius for the second closest set of nearest neighbors, or *n*=2

## Step 6. Reject outliers

The calculation to determine the boundaries for rejection of the outlier pixels is defined below in the equations and diagram.

**Assumptions for default value of 1.42**   The following assumptions lead to the default value of 1.42 for this parameter.

- Normal distribution for pixel intensity, where y-axis corresponds to pixel frequency and x-axis corresponds to pixel intensity.

- A 99% confidence interval that the pixels of interest are contained within the boundaries for rejection.

The *Interquartile Range (IQR)* is the range of points under a Gaussian distribution contained between the 25th percentile mark (25% of the points are contained under the curve from the zero point to the 25th percentile mark) and the 75th percentile mark. The 50th percentile mark is coincident with the median of the curve.

The *boundary for rejection* is the point on the x-axis beyond which all pixels will be rejected.

*"D"* is the distance between the mean of the curve and the boundary for rejection.

**Calculations of default value**   The following calculations are based on the above assumptions.

- If a pixel is located within the 99% confidence interval, it is 2.6 standard deviations (SD) away from the mean. Or, D = 2.6*SD and $D = Mult\_factor \times IQR + \kappa$.

- From the Z table for cumulative normal frequency distribution, the $Z_{P=0.75}$ = 0.675.

  Therefore,   $\kappa = 0.675 \times$ SD = IQR/2

- If you combine the four equations above and solve for the *Mult_factor*, the *Mult_factor* = 1.42.

- If you would rather use a 95% confidence interval, IQR *Mult_factor* = 0.952. The reason for this is, assuming normal distribution and infinite degrees of freedom, D = 1.96 * SD = $0.95185 \times IQR + \kappa$ .
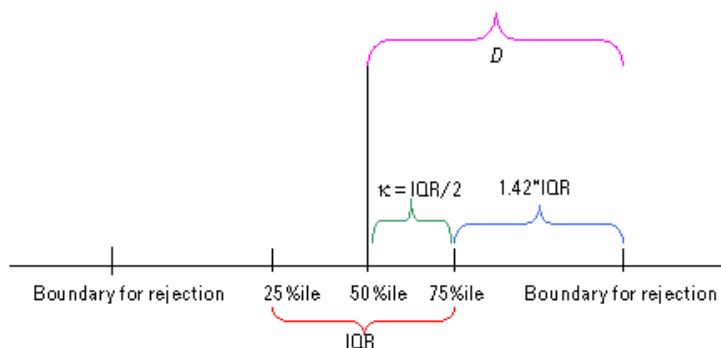


**Figure 45**   Important points on Gaussian curve—# of pixels vs. intensity

**Step 7. Calculate the mean signal of the feature (MeanSignal)**

The intensities of inlier pixels of a feature are averaged to give mean signal of the feature before background subtraction. The NumPix column in the result file lists the number of inlier pixels in the cookie that remain after rejection of outlier pixels.

$$MeanSignal = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \textbf{[4]}$$

where $n$ is the # of inlier pixels (i.e. NumPix), and $X_i$ is pixel intensity in the feature

The number of pixels that are removed as outliers at the high end and low end of the intensity distribution are shown in 4 columns of the FEATURES table: NumPixOLLo and NumPixOLHi (for both red and green channels).

## Step 8. Calculate the mean signal of the local background (BGMeanSignal)

The intensities of local background inlier pixels are averaged to give the local background mean signal. The BGNumPix column in the result file lists the number of inlier pixels in the local background radius that remain after rejection of outlier pixels.

$$BGMeanSignal = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \textbf{[5]}$$

where $n$ is the # of inlier pixels in the local background (i.e. BGNumPix), and $X_i$ is the pixel intensity in the local background

## Step 9. Determine if the feature is saturated (IsSaturated)

Feature is saturated if 50% of inlier pixels have intensity values above the saturation threshold.

# Flag Outliers

spacer

$\sigma^2_M$ is the *measured variance* of inlier pixels in the feature or background (e.g. PixSDev2 or BGPixSDev2).

$\sigma^2_E$ is the *estimated variance* using known noise characteristics of the Agilent Microarray Gene Expression system.

*For more information on confidence interval, check Numerical Recipes in C (Chapter 15, page 692).*

*Net signal* is the mean signal (i.e. MeanSignal or BGMeanSignal, respectively) minus the MinSigArray, which is minimum feature signal or minimum local background signal on the microarray, representing an estimate of the scanner offset.

## Step 10. Determine if the feature is a non-uniformity outlier (IsFeatNonUnifOL)

The non-uniformity outlier algorithm flags anomalous features and local backgrounds based on statistical deviations from the Agilent noise model. Feature or background is flagged as a non-uniformity outlier (e.g. IsFeatNonUnifOL or IsBGNonUnifOL, respectively) if the *measured variance* is greater than the product of the *estimated variance* and the *confidence interval multiplier*.

$$\sigma_M{}^2 > (\sigma_E{}^2 \times CI)$$ where *CI* is the confidence interval calculated from chi square distribution

The equations below are calculated for each feature and background per channel.

### Estimated Feature or Background Variance

The Agilent noise model estimates the expected variance by using noise effects from the Agilent Microarray Gene Expression system, which includes microarray manufacture, wet lab chemistry, and scanner noise.

$$\sigma^2{}_E = \sigma^2{}_{Labeling/FeatureSynthesis} + \sigma^2{}_{Counting} + \sigma^2{}_{Noise} \quad \textbf{[6]}$$

$$\sigma^2{}_E = Ax^2 + Bx + C \quad \textbf{[7]}$$

$x$ is the net signal of feature or background.

A or $\sigma^2{}_{Labeling/FeatureSynthesis}$ is the term that estimates the sources of variance that are proportional to the square of the signal, including microarray manufacturing and wet chemistry effects; the variance follows a Gaussian distribution. This term is intensity dependent and is the square of the CV (e.g. coefficient of variation) estimate of the pixel noise.

$$CV = \frac{PixSDev}{MeanSignal - MinSig_{Array}}[\mathbf{8}]$$

where B or $\sigma^2_{Counting}$ is the term that estimates the sources of variance that are proportional to the square-root of the signal, including scanning measurement or counting error; the variance follows a Poisson distribution. This term is dependent on the intensity and the scan resolution of the image.

where C or $\sigma^2_{Noise}$ is the term that estimates the sources of variance that are independent of the signal, including electronic noise in scanner and background level noise in glass; the variance is a Constant.

The variables A, B and C have different values for feature and background. For Agilent data produced with the GE2-SSPE_95_Feb07 protocol, these values are determined empirically (default selection in protocol) from self-vs-self experiments and from the known noise characteristics of the Agilent Microarray system discussed above. For all other Agilent Feature Extraction protocols, only the A term is empirically determined.

For all other Agilent protocols, the default selection in the protocol is to determine the B and C terms automatically. Here is how the Feature Extraction program calculates these terms:

- Saturated features are omitted from the population of negative control probes (NC). This NC set and the local background regions associated with these features are used in the calculations.

- Calculates Net Signal.

- Calculates the pixel standard deviation and then squares it to yield the pixel variance.

- From a histogram plot of number of features or bkgd vs. net signal, finds the net signal value for the 25th percentile.

- From a histogram plot of number of feature or local bkgd vs. variance, finds the variance for the 25th percentile.
- Calculates the B term as 25%NetSignal X B Term Multiplier and the C term as 25%Variance X C Term Multiplier.

For a given scanner, multipliers need to be determined. This tuning should use many images from different batches of microarrays, different users, and different processes. Different channels may need their own multipliers.

**Measured Feature or Background Variance**

$$\sigma_M^2 = \frac{1}{n-1} \times \sum_{i=0}^{n-1} (X_i - \overline{X})^2 \quad \textbf{[9]}$$

where $n$ is # of inlier pixels in the feature or background (i.e. NumPix or BGNumPix, respectively).

where $X_i$ is raw pixel intensity in the feature or background. (inlier pixels)

where $\overline{X}$ is mean raw pixel intensity for the feature or background (i.e. MeanSignal or BGMeanSignal, respectively).

## Step 11. Determine if the feature is a population outlier (IsFeatPopOL)

Agilent provides two different statistical algorithms for identifying population outliers. You select the appropriate algorithm to use in the protocol.

For probe sequences with enough replicate features, Feature Extraction uses the IQR test for population outlier analysis. The minimum number of replicates needed is set by the protocol field, "Minimum Population" and is set to 10 as the default for most Agilent protocols.

If the protocol choice, "Use Qtest for Small Populations?" is set to True, the Q-test method is used when a probe sequence has fewer than the minimum population number of features. The Q-test choice is set to True for Agilent's newer protocols.

### Qtest for replicate features < minimum population number

Q-test allows population outlier flagging for probe sequences from one less than the minimum population number down to 3.

This test is especially useful for NegC probes on CGH microarrays. Flagging features as population outliers is needed to accurately calculate NegCAvg and SD statistics.

This algorithm uses the following equation:

$Qi = |Xi - Xnearest| \backslash |Xmax - Xmin|$

Where $Xi$ = the intensity of a probe sequence;

$Xnearest$ = the intensity of the nearest probe sequence in intensity

$Xmax$ = the intensity of the most intense probe sequence

$Xmin$ = the intensity of the least intense probe sequence

Qi is compared to Qcritical to determine if the feature is an outlier. Qcritical depends upon the number of replicate features (N) and upon the chosen confidence level.

Agilent has chosen a 95% confidence level and bases the identification of population outliers on this table:

**Table 19** Qcritical values at 95% confidence level

| Number of replicated features (N) | Qcritical |
|---|---|
| 3 | 0.970 |
| 4 | 0.829 |

**Table 19**  Qcritical values at 95% confidence level (continued)

| Number of replicated features (N) | Qcritical |
|---|---|
| 5 | 0.710 |
| 6 | 0.625 |
| 7 | 0.568 |
| 8 | 0.526 |
| 9 | 0.493 |
| 10 | 0.466 |

**IQR Test for replicate features > or = minimum population number**

The equations below are calculated for each feature and background population per channel.

The intensities of all features or background regions in the population are plotted on a distribution curve. The difference in intensities between the $25^{th}$ and $75^{th}$ percentiles represent the Interquartile Range (IQR).
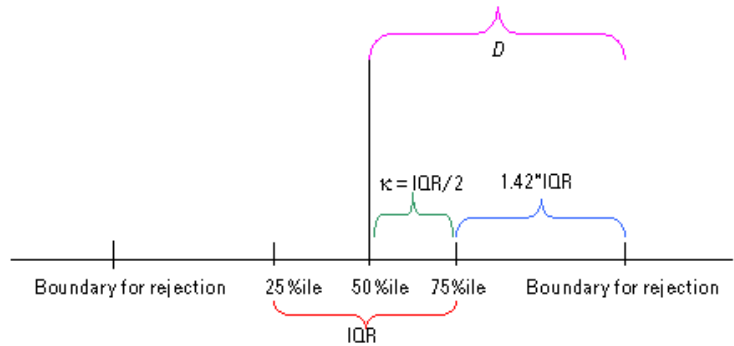
**Figure 46**   Interquartile Range

$$Cutoff_{PopOutlier} = 1.42 \times IQR \quad [10]$$

where $IQR$ = Intensity at $75^{\text{th}}$ percentile – Intensity at $25^{\text{th}}$ percentile.

where 1.42 is the IQR factor. Agilent uses 1.42 as the IQR factor so that the cutoff boundaries encompass 99% of the expected population distribution. The user can change this factor to encompass different boundaries, as discussed in the *Agilent Feature Extraction for CytoGenomics User Guide.*

Feature or background is flagged as population outlier (e.g. IsFeatPopOL or IsBGPopOL, respectively) if the mean signal (e.g. MeanSignal or BGMeanSignal) is greater than the upper rejection boundary (RBupper) or less than the lower rejection boundary (RBLower).

$MeanSignal > RB_{Upper}$

$MeanSignal < RB_{Lower}$

where

$RB_{Upper} = I_{75percentile} + Cutoff_{PopOutlier}$

and

$RB_{Upper} = I_{25percentile} - Cutoff_{PopOutlier}$

# Compute Bkgd, Bias and Error

Feature extraction completes several steps in order to determine the error model for each feature. First it determines and subtracts the background for each feature on the array. This is followed by detrending the array for systematic error. Finally an error model accounts for systematic and random errors encountered during sample preparation, hybridization, and scanning steps.

### Step 12. Calculate the feature background-subtracted signal (BGSubSignal)

The feature background-subtracted signal, BGSubSignal, is calculated by subtracting a value called the BGUsed from the feature mean signal.

$$BGSubSignal = MeanSignal - BGUsed \quad \text{[11]}$$

where BGSubSignal and BGUsed depend on the type of background method and the settings for spatial detrend and global background adjust. See the table below.

**Table 20** Values for BGSubSignal, BGUsed and BGSDUsed for different methods and settings[*]

| Background Subtraction Method | Background Subtraction Variable | Spatial Detrend (SpDe) OFF Global Bkgnd Adjust (GBA) OFF | SpDe ON GBA OFF | SpDe OFF GBA ON | Spatial Detrend ON Global Bkgnd Adjust ON |
|---|---|---|---|---|---|
| No background subtract | BGUsed = | BGMeanSignal[†] | SpatialDetrend SurfaceValue | BGAdjust | SpatialDetrendSurface Value (SDSV) + BGAdjust |
| | BGSDUsed = | BGPixSDev[‡] | BGPixSDev | BGPixSDev | BGPixSDev |
| | BGSubSignal = | MeanSignal | MeanSignal - BGUsed | MeanSignal - BGUsed | MeanSignal - BGUsed |
| Local Background | BGUsed = | BGMeanSignal | BGMeanSignal + SDSV | BGMeanSignal + BGAdjust | BGMeanSignal + SDSV + BGAdjust |
| | BGSDUsed = | BGPixSDev | BGPixSDev | BGPixSDev | BGPixSDev |

**Table 20** Values for BGSubSignal, BGUsed and BGSDUsed for different methods and settings[*] (continued)

| Background Subtraction Method | Background Subtraction Variable | Spatial Detrend (SpDe) OFF Global Bkgnd Adjust (GBA) OFF | SpDe ON GBA OFF | SpDe OFF GBA ON | Spatial Detrend ON Global Bkgnd Adjust ON |
|---|---|---|---|---|---|
| | BGSubSignal = | MeanSignal - BGUsed | MeanSignal - BGUsed | MeanSignal - BGUsed | MeanSignal - BGUsed |
| Global Background method | BGUsed = | GlobalBGInlierAve[**] (GBGIA) | GBGIA + SDSV | GBGIA + BGAdjust | GBGIA + SDSV + BGAdjust |
| | BGSDUsed = | GlobalBGInlierSDev (GBGISD) | GBGISD | GBGISD | GBGISD |
| | BGSubSignal = | MeanSignal - BGUsed | MeanSignal - BGUsed | MeanSignal - BGUsed | MeanSignal - BGUsed |

\* For both the red and green channels (2-color, CGH and non-Agilent microarrays)

† With No background subtraction as the setting, BGMeanSignal is the value for BGUsed only for the t-test, but no BGUsed is subtracted from the MeanSignal to produce BGSubSignal.

‡ If the method in the protocol for calculating the spot value from pixel statistics is Median/Normalized InterQuartile Range instead of Mean/Standard Deviation, the program makes these substitutions for the spot value and background subtraction calculations:
MedianSignal for MeanSignal
BGMedianSignal for BGMeanSignal
PixNorm IQR for PixSDev
GPixNormIQR for BGPixSDev
NormIQR = 0.7413 x IQR

\*\*If Median is the selection in the protocol, the median is substituted for the mean in the inlierAve and the InlierSDev calculations.

### Step 13. Perform background spatial detrending to fit a surface

To calculate the spatial shape or surface for each channel, the Feature Extraction program uses one of these background subtraction protocol selections:

• All Feature Types

This selection fits the surface to a set of very low intensity features evenly distributed on the slide using a "moving windowed filtering".

This algorithm, which was the original algorithm for gene expression microarrays, moves a window over the whole microarray and attempts to choose a fixed number of data points with the lowest intensity inside each window. This option is recommended for those arrays without negative controls and is illustrated in the following figure:
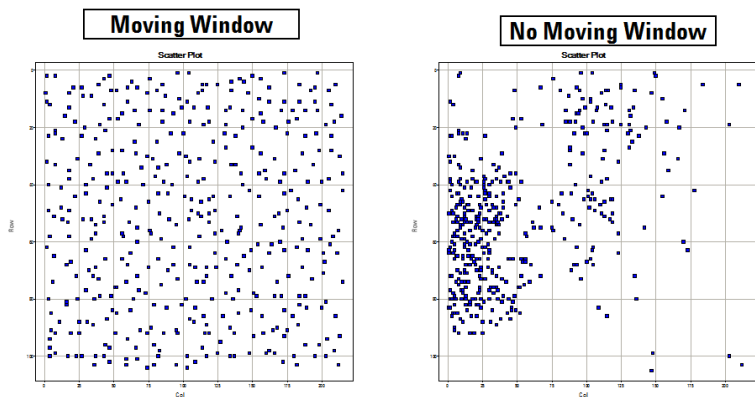


The effect of a moving window on selecting the lowest intensity features as an estimate of background. In the figures above, the blue squares represent the low intensity features found on the array. In the absence of a moving window, the lowest features on the entire array are located and may exhibit spatial bias. With the moving window, the lowest features from each region of the microarray are better identified.

• OnlyNegativeControlFeatures

This selection fits the surface to the negative control features distributed on the slide and is recommended for Agilent CGH microarrays.

This option works well with well defined negative controls. Outlier filtering should be enabled with this option to ensure good negative control values. To enable outlier filtering, set "NegCtrlSpread Outlier Rejection On" to True, which removes artifacts from distorting the control feature set distribution. This is illustrated in the following figure:
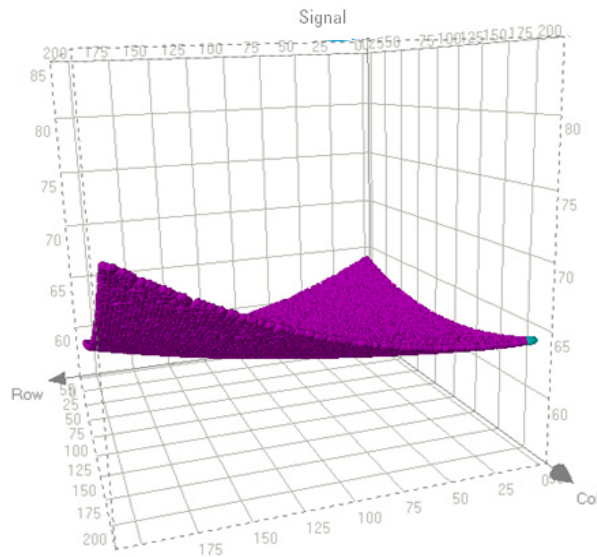


The purple surface represents a smoothed fit to all the negative control feature inliers. The residual of the surface fit is the Error on background subtraction in the Additive Error Estimation (see "Step 16. Determine the error in the signal calculation" on page 202).

- FeaturesInNegativeControlRange

  This algorithm does two levels of filtering. First, it finds the features in the range of negative controls, by fitting the negative controls to a surface and finding non-control features whose signal is within 3 standard deviations of that fit. Then, it fits a Lowess curve to this set of features. It interpolates from that fit to calculate a background signal for each feature.

  For high density microarrays, this algorithm can take a long time to complete its calculations. To speed up the process, you can elect in the protocol to randomly select a small percentage of the total points with which to calculate the fit. To do this, you set "Perform Filtering for Fit" to True, which significantly reduces the amount of time for spatial detrending of high density microarrays.
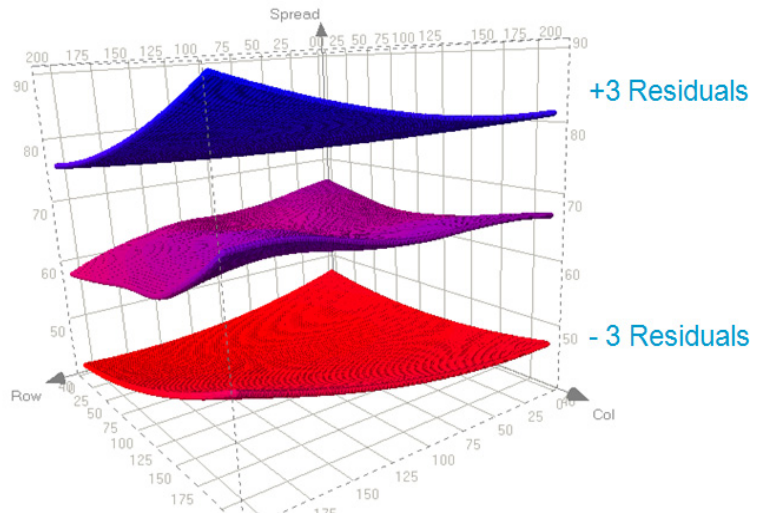


The purple surface represents the smoothed fit of all features, plus or minus 3 errors of the negative control fit. The residual of the surface fit is the Error on background subtraction in the Additive Error Estimation (see "Step 16. Determine the error in the signal calculation" on page 202).

The FeaturesInNegativeControlRange algorithm has been shown to more accurately estimate zero than the All Feature Types background algorithm. This improvement is shown below by viewing the features used in the additive detrend algorithm (colored in blue) superimposed on the InterpolatedNegCtrlSubSignal distribution. You can see that the signals of those features are closer to zero when the FeaturesInNegativeControlRange algorithm is used.
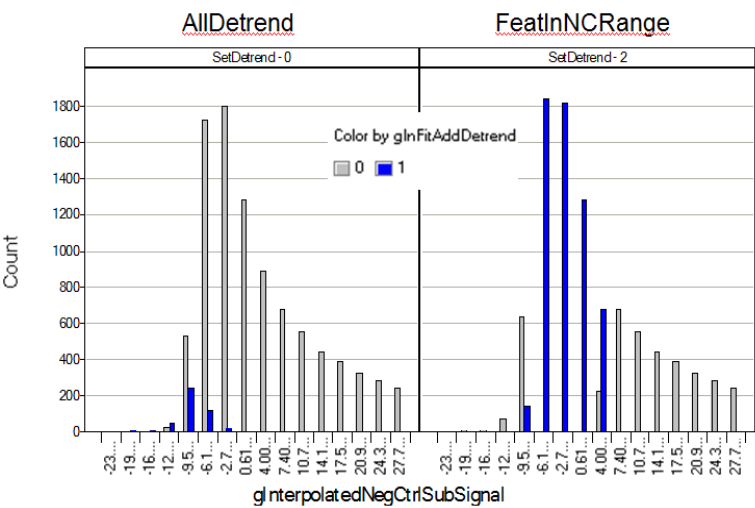


The effects of using all features for detrending (shown in the left figure) as compared to using the features in the negative control range (shown in the right figure). Features that had detrending added are shown in blue. The FeaturesInNegativeControlRange algorithm more accurately centers the values around zero.

A 2D-Loess algorithm fits the surface on the mean intensities of the filtered low intensity features of both red and green channels separately. This is described graphically in the figure below.
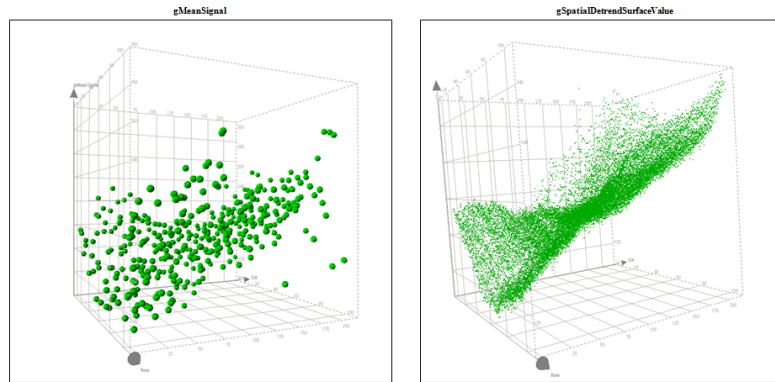
The effect of a 2-dimensional Loess fit to the green mean signal intensities across the array. You can find more information on the algorithm from the Web site
http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm

If $N$ = number of data points selected for surface fitting after filtering and $I_i$ = $i^{th}$ point from the filtered low intensity data set, the Loess algorithm fits a surface through these data points to obtain an intensity value describing the surface corresponding to each input data point.

Let $O_i$ denote the fitted output surface corresponding to the $i^{th}$ input point $I_i$. The statistical results that come out of this calculation are described in the table on the next page.

**Table 21**  Statistical results of spatial detrend algorithm

| Result | Description and Equation |
|---|---|
| SpatialDetrendRMSFit | This result gives an idea of the extent of the surface fit. It is the root mean square of the fitted data points obtained from the Loess algorithm. $$\sqrt{\frac{\sum_{i=1}^{N}\left(O_i - \frac{\sum_{i=1}^{N} O_i}{N}\right)^2}{N}} \quad [12]$$ |
| SpatialDetrendRMSFiltered minusFit | This result is the approximate residual from the surface fit. The deviations of the input (filtered) points from the corresponding output (fitted) data points are computed. An outlier rejection is performed on the set of deviations using the standard IQR technique (Figure 46 on page 188). Here $I$ is the value from the Loess fit and $O$ is the *BGSubSignal*. $$\sqrt{\frac{\sum_{i=1}^{N'}(I_i - O_i)^2}{N'}} \quad [13]$$ |
| SpatialDetrendSurfaceArea | This result gives an idea of the curvature of the surface gradient. |

**Table 21** Statistical results of spatial detrend algorithm (continued)

| Result | Description and Equation |
|---|---|
| SpatialDetrendVolume | The volume is calculated as the sum of the intensities of the surface area minus the offset. The offset is calculated as the volume under the flat surface (parallel to the glass slide) passing through the minimum intensity point of the fitted surface. This number (total volume - offset) is normalized by the area of the microarray. |
| SpatialDetrendAveFit | This describes the average intensity of the surface gradient. $$\frac{\sum_{i=1}^{N} O_i}{N} \quad [14]$$ |

## Step 14. Adjust the background

This algorithm determines the offset in both the red and green channels by identifying features that are not differentially expressed and fall within the central tendency of the data, especially in the lower intensity domain. These features should not be saturated or be flagged as non-uniform outliers.

Using this method yields more accurate and reproducible background-subtracted signals and log ratios for two-channel data than using no correction or single-channel correction.

Using a self-self microarray (i.e. same target labeled in red and green channels), one expects to see a linear plot of red background-subtracted signal versus green. If the backgrounds have not been estimated correctly in one channel with respect to the second channel, there will be a bias. This bias yields a "hook" at the low end of the signal range when shown in a plot with log scale axes (see Figure 47).
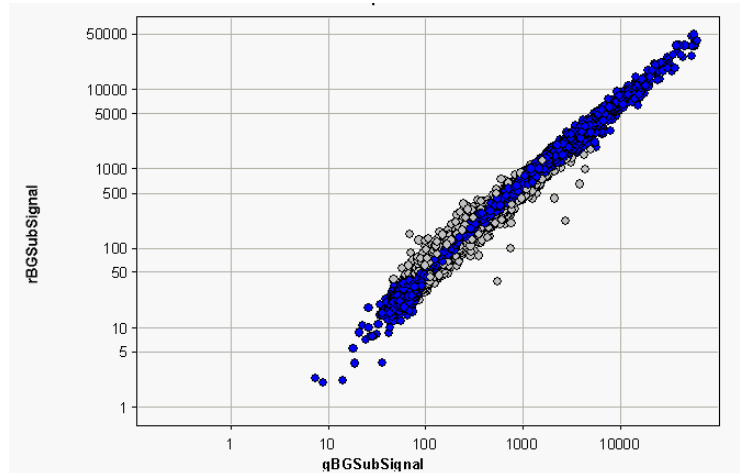
**Figure 47**    Unadjusted background-subtracted signals

The background adjustment algorithm first finds the central tendency of the data (features shown as blue circles in the figures). Using this subset of features, the algorithm then estimates the best adjustment in both the red and green channels to remove the bias. After the background adjustment, the bias is removed and the plot is linear (Figure 48).
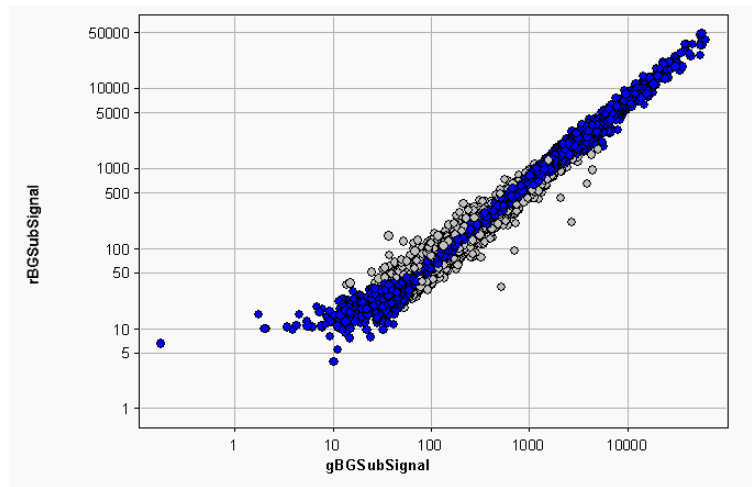
**Figure 48**    Adjusted background-subtracted signals

The bias, if uncorrected, yields a log ratio versus signal plot that is not symmetric about the log ratio axis (Figure 49); whereas, after adjustment, the data is more symmetric (Figure 50).
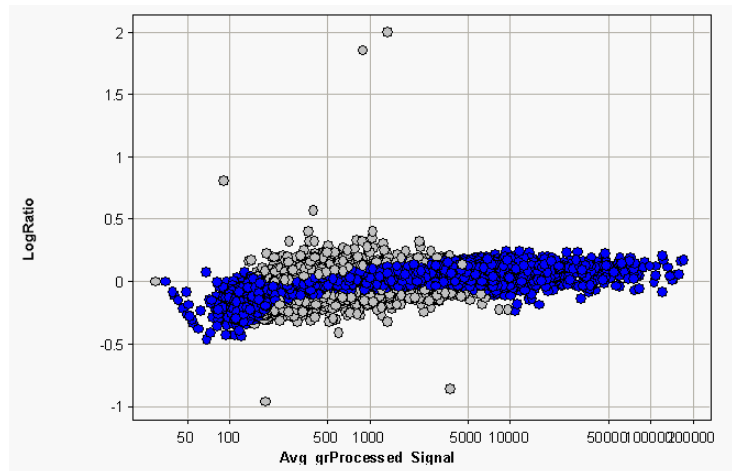


**Figure 49**    Log ratios calculated from unadjusted background-subtracted signals
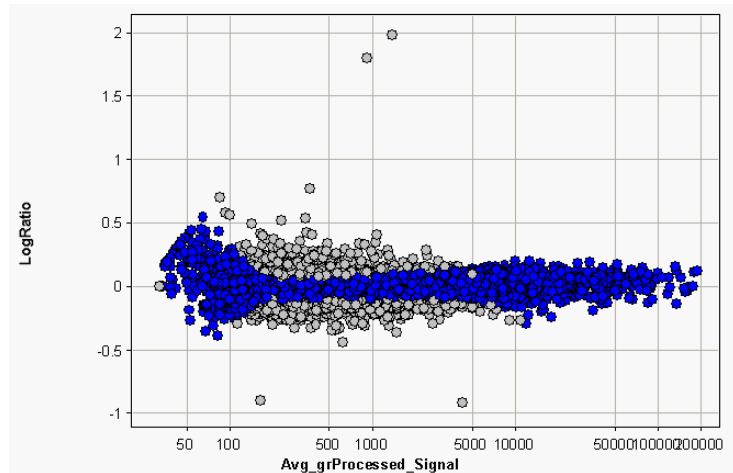
**Figure 50**    Log ratios calculated from adjusted background-subtracted signals

**How is the Adjust background globally "pad" used?**    If *Adjust background globally* is selected, you can enter a constant between 0 and 500, called the pad value, which forces the log ratio of red/green towards zero.

The value of the pad is expressed in raw counts, before dye normalization. The Feature Extraction program assumes that this value applies to the red or green channel with the smallest mean signal and automatically computes the corresponding raw value in the other channel that would yield a corrected log ratio of zero after dye normalization.

The red and green feature signals are analyzed for rank consistency. If red signal is plotted vs. green signal and the slope of the rank consistent features is >1, then the pad value is assigned to the green channel. If the slope is <1, the value is assigned to the red channel.

For instance, if you set *Adjust background globally to* 50, and if the slope is 1.2, then a value of 50 is added to the green background-subtracted signal of all features; whereas, a value of (50*1.2) = 60 is added to the red background-subtracted signal of all features.

Conversely, if you set *Adjust background globally to* 50, and if the slope is 0.5, then a value of 50 is added to the red background-subtracted signal of all features; whereas, a value of (50/ 0.5) = 100 is added to the green background-subtracted signal of all features.

## Step 15. Calculate robust negative control statistics

This algorithm repeats the population outlier algorithm, but not on one sequence at a time, rather on the distribution of all features that are classified as NegC or negative controls.

The algorithm calculates robust IQR statistics on features not designated as non-uniform outliers, population outliers or saturated.

UpperLimit = 75th percentile + Multiplier*IQR

LowerLimit = 25th percentile - Multiplier*IQR

The default value for this multiplier is 5.

The algorithm then omits features that are outside the Upper and LowerLimits and calculates the new robust Count, Avg, and SD of these inliers for the net signal and the background-subtracted signal:

g(r)NegCtrlNumInliers

g(r)NegCtrlAveNetSig

g(r)NegCtrlSDevNetSig

g(r)NegCtrlAveBGSubSig

g(r)NegCtrlSDevBGSubSig

## Step 16. Determine the error in the signal calculation

This step calculates the error on the background-subtracted and detrended signal. You can select for the error calculation either the Universal Error Model or the model (Universal or propagated) that produces the largest (most conservative) estimate of the error.

The Feature Extraction program does a dynamic computation of an approximation for the additive terms in both the red and green channels for the Universal Error Model. The estimation of the dynamic additive error term for each channel (red or green) is based on the following equation (for 1-color gene expression, the green channel):

$$AddError = \sqrt{m_1{}^2\sigma_{NegCtrl}{}^2 + m_2{}^2 DNF^2(RMSFit^2) + m_3{}^2 DNF^2(residual^2)} \quad \textbf{[15]}$$

where $m_1$ = *MultNcAutoEstimate*

$m_2$ = *MultRMSAutoEstimate*

$m_3$ = *MultResidualRMSAutoEstimate*

*DNF* = LinearDyeNormFactor of the corresponding channel

*residual* = The residual of the 2D Loess fit

Since the Additive Error is now calculated in Compute Background, Bias and Error Section, the DNF is 1 and the Variance of the NegCtrls are not scaled for the DNF either. This scaling is done to the AdditiveError after DyeNorm is completed.

$\sigma_{NegCtrl}{}^2$ = Variance of the inlier negative control

where inlier negative control implies the negative controls for the corresponding channel after rejections of saturated, population and non-uniform outliers.

where *SpatialDetrendRMSFit* = RMS of the points defining the surface fit for that channel. For more details on this term, see Table 21 on page 197.

For Agilent 8 x format oligo microarrays, the auto-estimation algorithm uses only the variance of the inlier negative controls. You can set *m1* or *m2* in equation 22 equal to zero in the protocol settings.

**MultNcAutoEstimate**  Multiplier for the first term in the additive error equation (standard deviation of the inlier negative control). The value changes depending on the protocol used:

GE1, GE2 and miRNA = 0

CGH and ChIP = 1

non-Agilent = 1

**MultRMSAutoEstimate**  Multiplier for the second term in the additive error equation (g(r)SpatialDetrendRMSFit). This term is proportional to the amount of sequence variability in the foreground.

On gene expression arrays, Agilent uses this term because there is a single sequence for all negative controls so an estimation of any sequence-dependent foreground noise using negative controls is not possible.

For CGH microarrays, the error model choice is to make this term and *m3* zero and use only *m1* because there are a variety of sequences used for the negative controls.

GE1, GE2 and miRNA = 0

CGH and ChIP = 0

non-Agilent = 4

**MultResidualRMSAutoEstimate**  Multiplier for the third term in the equation and is the width of the distribution of signals used in the background spatial detrending set (after the background surface has been subtracted out).

When the background detrending set includes a group of features well-distributed across the microarray with a variety of sequences, the width of the distribution of the signals of these features after background subtraction is a very good estimate of the uncertainty of the dim signals, or the additive error.

GE1, GE2 and miRNA = 1

CGH and ChIP = 0

non-Agilent = 0

## Step 17. Calculate the significance of feature intensity relative to background (IsPosAndSignif)

The significance of the feature intensity compared to the background intensity (local or global) is calculated using two different significance tests: one using pixel statistics for both the feature and the background values and the other using the additive error from the Error Model calculation for the background value.

**Significance based on pixel statistics**   This method to determine significance uses the 2-sided Student's t-test with mean signal for the feature and the background correction for the background. This is implemented as an incomplete Beta Function approximation.

$$t = \frac{(\overline{X}_F - \overline{X}_B)}{\sqrt{\frac{(n_F - 1)\sigma_F^2 + (n_B - 1)\sigma_B^2}{df} \times \left(\frac{1}{n_F} + \frac{1}{n_B}\right)}} \quad \text{[16]}$$

where $\overline{X}_F$ is the mean signal (MeanSignal) of the feature and $\overline{X}_B$ is the background correction used for subtraction (BGUsed − see Table 20 on page 190).

where $n_F$ and $n_B$ are the number of inlier pixels in the feature or background (local), respectively (e.g. NumPix or BGNumPix).

where $\sigma_F^2$ and $\sigma_B^2$ are variances of inlier pixels for feature and background, respectively (e.g. PixSDev$^2$ or BGSDUsed$^2$).

$$\sigma_F^2 = \frac{1}{n_F - 1} \sum_{i=0}^{n-1} (X_i - \overline{X}_F)^2 \quad \text{[17]} \quad X_i \text{ is pixel intensity}$$

$$\sigma_B^2 = \frac{1}{n_B - 1} \sum_{i=0}^{n-1} (X_i - \overline{X}_B)^2 \quad \text{[18]}$$

where $df$ is the degrees of freedom,

$$df = n_F + n_B - 2$$

After the p-value is calculated from the 2-sided t-test using incomplete Beta Function, it is compared to the user-defined max p-value. If the calculated p-value from the Beta Function is less than the user-defined max p-value, then the feature signal is considered to be significantly different from the background signal.

If $p\text{-}value_{Calculated} < p\text{-}value_{Max}$, and if MeanSignal > BGUsed, then feature gets a Boolean flag of 1 under the IsPosAndSignif column in Feature Extraction result file.

**Significance based on additive error**    The Error model significance also uses a Gaussian probability distribution for the calculation and tests to see if a signal is greater than 0 with a known additive error. We compute the probability in a similar way to the Pixel Significance calculation. But instead of having a feature signal and a background signal, the test uses the feature signal and one error (background signal distribution is assumed to be around 0 with one error).

The degrees of freedom are large enough to make the function Gaussian. We define the error as one standard deviation (1SD) from the probability of 0 on the Gaussian curve and equal to a p-value of .01 (AdditiveError/2.6).

If the probability is greater than or equal to 1SD or .01, the background-subtracted signal is flagged as positive and significant. If it is less than 1SD or .01, it is flagged as not significant.

The value of the surrogate is scaled by the probability returned. The surrogate value for the Not significant signals equals AddError/2.6 * the probability, calculated this way for two reasons.

- Signals stay continuous.
- Surrogate values are not larger than the smallest significant signals.

### Step 18. Determine if the feature background-subtracted signal is well above the background (IsWellAboveBG)

The feature background-subtracted signal (i.e. BGSubSignal) is compared to the noise of its background (local or global):

$$BGSubSignal > WellAboveSDMulti \ \text{x} \ SD_{BG}$$

where

*WellABoveSDMulti* is the well above SD multiplier (5, default) - this means a feature is well above background if its signal is 5 times the additive error.

$SD_{BG}$ is the background standard deviation (i.e. BGSDUsed)

For the Error model significance test, the SD becomes AddError/2.6.

If the background-subtracted signal is greater than the *WellAboveSDMulti* x $SD_{BG}$, and if the feature passes the IsPosAndSignif test, then the feature gets a Boolean flag of 1 under the IsWellAboveBG column in Feature Extraction result file.

### Step 19. Calculate the surrogate value (SurrogateUsed)

The surrogate value is calculated and used as the "lowest limit of detection" to replace the dye-normalized signal when any of the following situations occur. These tests are done for each channel:

- MeanSignal is less than BGUsed or not significant compared to BGUsed (i.e., IsPosAndSignif = 0).
- BGSubSignal is less than its background standard deviation (i.e., BGSubSignal < BGSDUsed).

  The decision to replace a dye-normalized signal with a surrogate value is not made, however, until after probes are selected for correcting the dye bias.

The surrogate value is calculated in this step using these criteria:

If pixel significance is used to calculate IsPosAndSignif, then

$$SurrogateUsed \ = \ SD_{BG} \qquad\qquad \textbf{[19]}$$

where $SD_{BG}$ is the background standard deviation (i.e. BGSDUsed)

For the local background method, the standard deviation of the background is at the pixel-level of the local background.

For global background methods, the standard deviation of the background is at the replicate background-population level of the microarray.

If Error model significance is used to calculate IsPosAndSignif, then

$$SurrogateUsed \ = \ AddError/LinearDyeNormFactor \ \ \textbf{[20]}$$

where AddError is the additive error from the Error Model calculation

If Multiplicative Detrending is used, the SurrogateUsed is scaled by the MultDetrendSignal for each feature.

If a p-value other than default 0.01 is chosen in the protocol, then the SurrogateUsed is adjusted appropriately.

### Step 20. Perform multiplicative detrending

Multiplicative detrending is an algorithm designed to compensate for slight linear variations in intensities that can occur if the processing is not homogeneous across the slide. This non-homogeneous processing results in different chemical reaction times, for example, between the sides and the center, and produces a "dome effect".

With 2-color microarrays these dome effects are the same in each channel and for the most part cancel out during the calculations. Agilent has found multiplicative detrending to still be useful, however, for all the microarrays.

This algorithm is designed to correct the data by fitting a smoothed surface via a second degree polynomial fit to the

higher signals on the microarray (after outliers are rejected). This is shown in the illustration below:
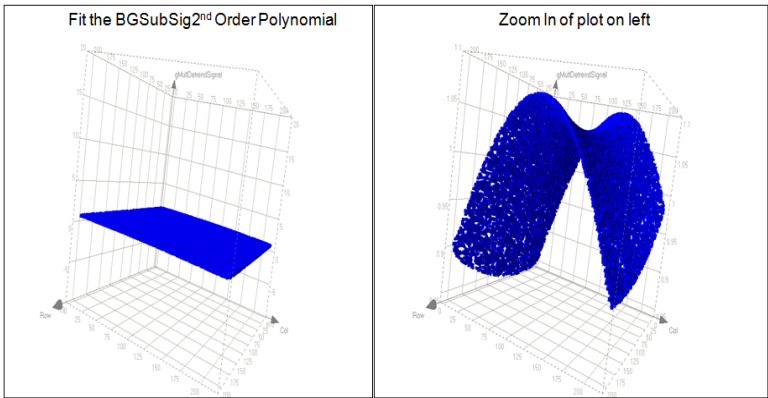


The effect of multiplicative detrending across array features. A second-order polynomial is fit to the higher signals on the array resulting in a subtle shape fit. This fit results in the ProcessedSignal having a better fit to the data than the BGSubSignal.

Because the multiplicative trend can be confused with the additive trend for dim microarrays, data points inside a multiple times the standard deviation from the center of the signals for the negative control population are excluded.

The equations for statistics and results that are produced by this calculation are shown in the following table. See Table 18, "Algorithms (Protocol Steps) and the results they produce," on page 166 for descriptions of these results.

**Table 22** Statistics and Results for Multiplicative Detrending

| Results | Equation |
|---|---|
| gMultDetrendRMSFit<br><br>*MDS =*<br>*MultDetrendSignal* | $$\sqrt{\frac{\sum\limits_{i=1}^{N}(MDS_i - average(MDS))^2}{N}}$$<br><br>[21] |
| gMultDetrendSignal | $$\frac{10^{Fitted(\log 10(BgSubSignal))}}{\left[\dfrac{\sum\limits_{i=1}^{N}(10^{Fitted(\log 10(BgSubSignal))})_i}{N}\right]}$$<br><br>[22] |
| gProcessedSignal | $$\frac{BGSubSignal_i}{MultDetrendSignal_i}\ [23]$$ |
| gProcessedSigError | $$\frac{BGSubSignalError_i}{MultDetrendSignal_i}\ [24]$$ |

# Correct Dye Biases

### Step 21. Determine normalization features

Normalization features are features used to evaluate the dye bias between the red and green channels.

**Using "All Probes" method**   Under this method, the initial normalization features are selected based on the following three criteria:

- Features are positive and significant versus the background (e.g. IsPosAndSignif = 1)

- Features are non-control (e.g. ControlType = 0)

- Features are non-outlier (e.g. IsFeatNonUnifOL = 0, IsFeatPopnOL = 0, IsSaturated = 0)

**Using "List of Normalization Genes" method**   Under this method, the user selects the normalization features. These features can be housekeeping genes or genes with no differential expression.

**Using "Rank Consistency Probes" method**   Under this method, the chosen normalization features simulate housekeeping genes. These features fall within the central tendency of the data, having consistent trends between the red and green channels. They are selected based on the following two criteria:

- Features pass the three criteria described in the "all significant, non-control, and non-outlier features" method and

- Features pass the rank consistency filter between the red and green channels

Rank consistency filter is done by transforming the feature BGSubSignal to feature rank per channel. Next, the feature correlation strength is calculated per feature:

$$CS = \frac{|\rho_R - \rho_G|}{N} \quad \textbf{[25]}$$

where $\rho_R$ and $\rho_G$ are the ranks of feature in the red and green channels, respectively

where $N$ is the total number of initial normalization features

If the $CS \leq \tau$, where $\tau$ is the threshold percentile, then feature passes the rank consistency filter between the red and green channels and falls within the central tendency of the data. Note $\tau$ is a user-defined parameter in the Feature Extraction program.

**Using "Rank Consistent List of Normalization Genes"** This method uses the rank consistent normalization genes from the list. These genes follow the criteria described above.

### Step 22. Calculate the normalization factor

**LinearDyeNormFactor** The linear dye normalization method assumes that dye bias is not intensity-dependent and therefore takes a global approach to dye normalization. A linear dye normalization factor is computed per channel by setting the geometric mean of signal intensity of the normalization features equal to 1000:

$$LinearDyeNormFactor = \frac{1000}{10^{\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right)}} \quad \textbf{[26]}$$

*The LinearDyeNormFactor (red and green channels) values are listed in the STATS table.*

where $X_i$ is the background-subtracted signal of a feature (i.e. BGSubSignal)

where $n$ is the number of features used for normalization (i.e. features with IsNormalization = 1)

**LOWESSDyeNormFactor** The LOWESS dye normalization method assumes that dye bias may be intensity-dependent and therefore takes a local approach to dye normalization.

The LOWESS dye normalization factor is calculated by fitting the locally weighted linear regression curve to the chosen normalization features. The amount of dye bias is determined from the curve at each feature's intensity. Each feature gets a different LOWESS dye normalization factor per channel.

The LOWESS method corrects the log ratio data so that its central tendency after dye normalization lies along zero for all intensity ranges, assuming an equal number of up- and down-regulated features in any given signal range. The

LOWESS DyeNormFactor is derived for each channel by the procedure described on the next page:

a A linear regression curve is fit to the data in a plot of M vs. A, where M (y axis) = Log(R/G) and A (x axis) = 1/2 x Log(R*G). R and G represent the red and green background-subtracted signals. This LOWESS curve fit through the central tendency of the M vs. A plot is defined as Mfit, and is a function of A.

b The dye normalization step transforms the data so that the central tendency of Mfit at every A is shifted to be equal to zero.

c After the correction factor is determined for any feature, it is split evenly over the red and green channels.

The new signals after correction, R' and G', are obtained by transforming the original R and G:

R' = R/($10^{MFit/2}$) and G' = G*($10^{MFit/2}$)

d If the original log ratio is exactly along the fit line Mfit, the new log ratio is shifted to zero:

If log(R/G) = Mfit, then Log(R) = Log(G) + Mfit

or Log(R'*$10^{MFit/2}$) = Log (G'*$10^{-MFit/2}$) + Mfit

or Log(R') + Mfit/2 = Log(G') - Mfit/2 + Mfit

or Log(R'/G') = 0

e The LOWESSDyeNormFactor for R is 1/($10^{M'/2}$).
The LOWESSDyeNormFactor for G is $10^{M'/2}$.

**Linear&LOWESSDyeNormFactor**  This curve fitting algorithm does a linear scaling/normalization of the data individually in each channel before performing a non-linear dye normalization.

*Note that the Linear&LOWESS dye normalization factor is not reported in the Feature Extraction output file. Therefore, the only way to know the Linear & Lowess dye norm factor is to calculate it using the equation below.*

The Linear&LOWESS dye normalization factor can be calculated from the equation below:

$$\text{Linear\&LOWESSDyeNormFactor} = \frac{DyeNormalSignal}{BGSubSignal \times LinearDyeNormFactor} \qquad \textbf{[27]}$$

### Step 23. Determine if surrogate values must substitute for low-intensity signals

At this point two criteria are used to determine is surrogate values must take the place of the low-intensity signals:

- The feature signal is not positive and significant versus background.
- The signal is not larger than the background error.

Surrogate values were computed during background subtraction and are stored in the SurrogateUsed column.

### Step 24. Calculate the dye-normalized signal (DyeNormSignal)

The dye-normalized signal is calculated by multiplying the background-subtracted signal by the dye normalization factor:

$$DyeNormSignal = BGSubSignal \times DNF \quad \text{[28]}$$

where *DNF = LinearDyeNormFactor*, when linear dye normalization method is used and where:

$$DNF = LinearDyeNormFactor \times LOWESSDyeNormFactor \quad \text{[29]}$$

when LOWESS dye normalization method is used.

# Compute Ratios

### Step 25. Calculate the processed signal (ProcessedSignal)

The processed signal is used in calculating the log ratio. If a surrogate is not used (i.e. SurrogateUsed = zero value), then the processed signal is the dye-normalized signal. If a surrogate is used (i.e. SurrogateUsed = non-zero value), then the processed signal is the SurrogateUsed value.

if *SurrogateUsed* = 0, then
*ProcessedSignal = DyeNormSignal*

if *SurrogateUsed* ≠ 0, then
*ProcessedSignal = SurrogateUsed * DyeNormFactors,*
*where DyeNormFactors = LinearDyeNormFactor ** 
*LowessDyeNormFactor, if Linear and Lowess methods are used*

### Step 26. Calculate the log ratio of feature (LogRatio)

The log ratio $i$ is the measure of differential expression between the red and green channels for every probe $i$:

$$LogRatio_i = Log_{10}\left(\frac{ProcessedSignal_{r,i}}{ProcessedSignal_{g,i}}\right) \quad \textbf{[30]}$$

where *ProcessedSignal*$_{r,i}$ and *ProcessedSignal*$_{g,i}$ are signals post dye normalization and post surrogate processing in the red and green channels, respectively.

### Step 27. Calculate the p-value and error on log ratio of feature (PvalueLogRatio and LogRatioError)

PvalueLogRatio gives the statistical significance on the log ratio per each feature (e.g. gene) between the red and green channels. The p-value is a measure of the confidence (viewed as a probability) that the feature is not differentially expressed.

For example, if the p-value is less than 0.01, we can say with a 99% confidence level that the gene is differentially

expressed. In other words, there would be a 1% random chance of getting this low of a p-value with a gene that is actually not differentially expressed:

$$\text{p-value} = 1 - Erf\left(\frac{|xdev|}{\sqrt{2}}\right) = Erfc\left(\frac{|xdev|}{\sqrt{2}}\right) \quad \text{[31]}$$

where:

$$Erf(x) = \frac{2}{\sqrt{pi}}\int_0^x e^{-t^2} dt \quad \text{[32]}$$

*Erf(x)* is the error function of the expression $x$ as given by the above equation: It is twice the integral of the Gaussian distribution with mean = 0 and variance = 1/2

*Erfc* is the complementary error function as defined by the above equation.

*xdev* is the deviation of LogRatio from 0.

$$xdev = \frac{LogRatio}{LogRatioError} \quad \text{[33]}$$

Equation 22 is analogous to a signal to noise metric.

*For more details on calculations with the Universal Error Model, see the confidential Agilent technical paper on error modeling.*

If the **Universal Error Model** is used, then xdev is computed from six sources:

• ProcessedSignals (red and green channels)
• Multiplicative error factors (red and green)
• Additive error factors (red and green)

The terms xdev, 'multiplicative error', and 'additive error' come from the Universal Error Model, as developed by Rosetta Biosoftware.

Once xdev is computed, it is plugged back into Equation 2, where LogRatioError is derived.

*For more details on calculations with the propagation error model, see the confidential Agilent technical paper on error modeling.*

If the **Propagation of Pixel Level Error Model** is used, then LogRatioError is computed from the following sources:

- Feature PixSDev (red and green channels)
- Background Noise (calculation is dependent upon the chosen BkSubMethod; red and green channels)

Once the LogRatioError is computed, it is plugged back into Equation 21, where xdev is derived.

**Table 23** Summary – Use of surrogates for calculations

| | |
|---|---|
| Case 1: R/G<br>Both channels use DyeNorm Signals.<br>P-value and log ratio are calculated as usual.<br>For signals not using surrogates,<br>g(r)DyeNormSignal = g(r)ProcessedSignal,<br>which is then used to calculate log ratio. | Case 2: r/G<br>r = rSurrogateUsed<br>G = gDyeNormSignal<br>P-value and log ratio are calculated as usual.<br>If r/G > 1, then Feature Extraction automatically sets<br>LogRatio = 0 and PvalueLogRatio = 1 |
| Case 3: R/g<br>R = DyeNormSignal<br>g = gSurrogateUsed<br>P-value and log ratio are calculated as usual.<br>If R/g < 1, then Feature Extraction automatically sets<br>LogRatio = 0 and pValueLogRatio = 1 | Case 4: r/g<br>Both channels use surrogates.<br>Feature Extraction automatically sets<br>LogRatio = 0 and pValueLogRatio = 1<br>For signals using surrogates,<br>g(r)ProcessedSignal =<br>g(r) SurrogateUsed * g(r)DyeNormFactors. |

# Calculate Metrics

Although the QC metrics are calculated in this step, only the gridding tests are discussed in this section.

### Step 28. Perform a series of gridding tests to make sure that grid placement has been successful

These tests are performed to yield warnings on the Summary Reports about unsuccessful gridding. They also produce the assessment shown in the QC Report of whether the grid needs to be evaluated or not.

In Feature Extraction, new tests have been added and thresholds tuned to decrease the number of false negatives

(Summary Report shows no problems when there are) and false positives (Summary Report shows a problem when there isn't).

The parameters for these tests do not appear in the protocols, but they do appear in the FEParams output.

Below is a question asked by each test, the metric used to answer the question ("stat" name that appears in the result text file as the Statistics table) and the threshold to assess gridding success or failure. If a grid fails any one of these tests, a warning or warnings appear in the reports.

**Test 1**   How many features are "not found" along the edge of the microarray?

Stat name: MaxSpotNotFoundEdges

Threshold_Max: 0.72

**Test 2**   How many local background regions are flagged as non-uniform outliers in either channel?

Stat name: AnyColorPrcntBGNonUnifOL

Threshold_Max: 2%

**Test 3**   How broad is the distribution of NegControl net signals?

Stat name: Max{gNegCtrlSDevNetSig, rNegCtrlSDevNetSig}

Threshold_Max: 100

**Test 4**   What is the median CV% of BGSubSignal of the NonControl replicated sequences?

Stat names: Max{gNegCtrlMedPrcntCVBGSubSig, rNegCtrlMedPrcntCVBGSubSig} or just the green stat for a 1-color application

Threshold_Max: 50%

**Test 5**   What is the difference between feature centers found by the gridding algorithm vs. the spot-finding algorithm?

Stat names: Max{CentroidDiffX, CentroidDiffY}

Threshold_Max: 10%

**Optional Test 6**     How many features along the edge of the microarray are flagged as non-uniform outliers in either channel?

This test is used only if one of these two metrics is unavailable:

- No replicated features are present to calculate the NonCtrlMedPrcntCVBGSubSig metric.
- Or no NegControls are present to calculate the StdDev.

Stat name: MaxNonUnifEdges

Threshold_Max: 10%

# Example calculations for feature 12519 of Agilent Human 22K image



Feature: 12519, Found
Gene name: DRD4
Intensity Red: 13502.5166, Green: 3021.7742
Red/Green Log Ratio: 0.0309
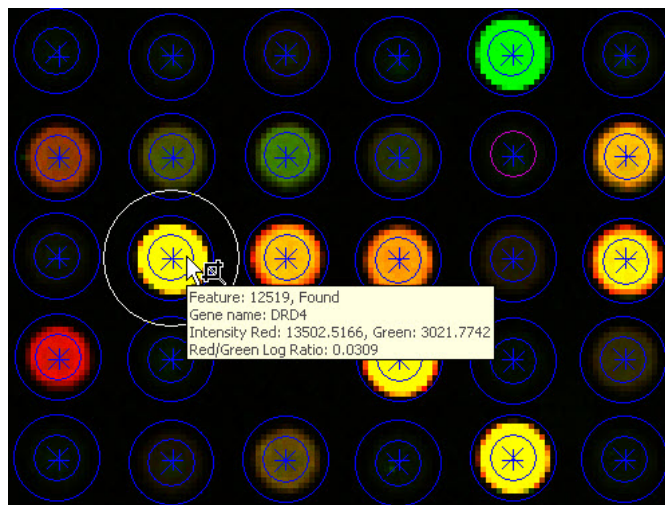
**Figure 51**   Visual results of feature number 12519 from "Shapes" file (*.shp) of Human_22K_expression microarray image

The 2-color gene expression Human 22K microarray image, "Human_22K_expression", is included in the Example Images that Agilent provides on the Feature Extraction software installation CD.

# Data from the FEPARAMS table

| BGSubtractor_BGSubMethod | BGSubtractor_BackgroundCorrectionOn | BGSubtractor_SpatialDetrendOn |
|---|---|---|
| 7 | 0 | 1 |

The BGSubMethod of 7 corresponds to No Background Subtraction method (see Table 3 on page 71 of this guide.). Global Background Adjustment is turned Off. Spatial Detrending is turned On.

# Data from the STATS Table

*LowessDyeNormFactor is not shown in Feature Extraction result file. This value can be back calculated using DyeNormSignal equation on page 245.*

| gLinearDyeNormFactor | rLinearDyeNormFactor |
|---|---|
| 15.881 | 4.14607 |

# Data from the FEATURES Table

## Results from Find And Measure Spots Algorithm

| FeatureNum | gNumPix | rNumPix | gMeanSignal | rMeanSignal | gPixSDev | rPixSDev |
|---|---|---|---|---|---|---|
| 12519 | 62 | 62 | 3021.774 | 13502.52 | 187.8805 | 1102.547 |

**Results from Correct Bkgd and Signal Biases Algorithm**

| FeatureNum | gSpatialDetrendSurfaceValue | rSpatialDetrendSurfaceValue |
|---|---|---|
| 12519 | 81.5464 | 72.2993 |

| FeatureNum | gBGUsed | rBGUsed | gBGSDUsed | rBGSDUsed | gBGSubSignal | rBGSubSignal |
|---|---|---|---|---|---|---|
| 12519 | 81.5464 | 72.2993 | 3.5514 | 5.34552 | 2940.23 | 13430.2 |

| FeatureNum | gIsPosAndSignif | rIsPosAndSignif | gIsWellAboveBG | rIsWellAboveBG |
|---|---|---|---|---|
| 12519 | 1 | 1 | 1 | 1 |

$$rBGUsed \ = \ rSpatialDetrendSurfaceValue$$

$$\textbf{72.2993 = 72.2993}$$

Note that this equation is valid only if there is no background subtraction, spatial detrending is on, and there is no global background adjustment.

*For an explanation of BGUsed with other background settings, see Table 20 on page 190.*

$$rBGSubSignal \ = \ rMeanSignal \ - \ rGBGUsed$$

$$\textbf{13430.2 = 13502.52 - 72.2993}$$

**Results from Correct Dye Biases Algorithm**

*Refer to "Data from the STATS Table" on page 221 for the LinearDyeNormFactor value.*

| FeatureNum | gDyeNormSignal | rDyeNormSignal |
|---|---|---|
| 12519 | 45834.1 | 49209.6 |

$$rDyeNormSignal \ = \ rBGSubSignal \ x \ rLinearDyeNormFactor \ x \ rLOWESSDyeNormFactor$$

$$\textbf{49209.6 = 13430.2 x 4.14607 x } rLOWESSDyeNormFactor$$

## Results from Compute Ratios and Errors Algorithm

| FeatureNum | gSurrogateUsed | rSurrogateUsed | gProcessedSignal | rProcessedSignal |
|---|---|---|---|---|
| 12519 | 0 | 0 | 45834.13 | 49209.64 |

| FeatureNum | LogRatio | LogRatioError | PValueLogRatio |
|---|---|---|---|
| 12519 | 0.0308611696 | 0.06148592089 | 0.6157220099 |

For the red channel, does the feature number 12519 pass the two criteria listed below that are required to calculate an accurate and reproducible log ratio?

- Feature is positive and significant vs. background (i.e. IsPosAndSignif = 1.

- BGSubSignal is greater than its background standard deviation (i.e. BGSDUsed).

For this example calculation, feature number 12519 passed both criteria. Since rSurrogateUsed = 0, the rDyeNormSignal is the same value as the rProcessedSignal.

$$rProcessedSignal = rDyeNormSignal, \ if \ rSurrogateUsed =$$
$$49209.6 = 49209.6$$

If a feature fails either or both of the criteria above, SurrogateUsed is a non-zero value and is calculated as shown below, depending on the Significance test parameter chosen in the **Compute Bkgd, Bias, and Error** protocol step.

> *rSurrogateUsed = rAddError/rLinearDyeNormFactor*
> *if Use Error Model for Significance is selected*

*rSurrogateUsed = rBGSDUsed*
*if Use Pixel Statistics for*
*Significance is selected*

If a surrogate is used in the red channel (i.e. rSurrogateUsed is a non-zero value), the red processed signal is calculated as surrogate value multiplied by the dye normalization factors.

> *rProcessedSignal = rSurrogateUsed * rLinearDyeNormFactor **
> *rLowessDyeNormFator*, if *rSurrogateUsed* $\neq 0$

The Log ratio is the log of red processed signal over green processed signal.

$$LogRatio = \log\frac{rProcessedSignal}{gProcessedSignal}$$

**0.0308612 = log (49209.64 / 45834.13)**

It is important to note that log ratio and p-value calculations are computed differently, depending on whether a surrogate is used in only one channel, both channels, or neither channels.

If a feature uses a surrogate in only the red channel (Case 2 of Table 23) and the red surrogate value is not greater than the green processed signal, the p-value and error on the log ratio are calculated, as usual, using equations 1 and 2 in this guide.

# Index

**www.agilent.com**

## In This Book

The *Reference Guide* presents descriptions of the protocols, or methods, available for use with Agilent Feature Extraction for CytoGenomics, as well as a listing of results and an explanation of how the Feature Extraction algorithms work.

This guide provides:

- a list of the default settings for each protocol shipped or downloaded with the software

- a list of all the parameters and results available after feature extraction

- the equations and a sample calculation for the feature extraction process

**Agilent Technologies**