

# arrayQualityMetrics report for QA0112F

- [Section 1: Between array comparison](#)
  - Distances between arrays
  - Principal Component Analysis
- [Section 2: Array intensity distributions](#)
  - Boxplots
  - Density plots
- [Section 3: Variance mean dependence](#)
  - Standard deviation versus rank of the mean
- [Section 4: Affymetrix specific plots](#)
  - Relative Log Expression (RLE)
  - Normalized Unscaled Standard Error (NUSE)
  - RNA digestion plot
- [Section 5: Individual array quality](#)
  - MA plots
  - Spatial distribution of M

## - Array metadata and outlier detection overview

array	sampleNames	*1	*2	*3	*4	*5	*6	sample	ScanDate
<input type="checkbox"/>	1 0112F-02_A18_(MoGene-1_0-st-v1).CEL							1	2012-02-15T20:57:03Z
<input type="checkbox"/>	2 0112F-02_A28_(MoGene-1_0-st-v1).CEL							2	2012-02-15T21:05:56Z
<input type="checkbox"/>	3 0112F-02_A32_(MoGene-1_0-st-v1).CEL							3	2012-02-15T22:16:04Z
<input type="checkbox"/>	4 0112F-02_A36_(MoGene-1_0-st-v1).CEL							4	2012-02-15T22:24:58Z
<input checked="" type="checkbox"/>	5 0112F-02_A44_(MoGene-1_0-st-v1).CEL				x			5	2012-02-15T20:39:17Z
<input type="checkbox"/>	6 0112F-02_F52_(MoGene-1_0-st-v1).CEL							6	2012-02-15T20:03:38Z
<input type="checkbox"/>	7 0112F-02_F58_(MoGene-1_0-st-v1).CEL							7	2012-02-15T20:12:27Z
<input type="checkbox"/>	8 0112F-02_F60_(MoGene-1_0-st-v1).CEL							8	2012-02-15T21:23:48Z
<input type="checkbox"/>	9 0112F-02_F64_(MoGene-1_0-st-v1).CEL							9	2012-02-15T21:14:57Z
<input type="checkbox"/>	10 0112F-02_F66_(MoGene-1_0-st-v1).CEL							10	2012-02-15T21:32:42Z
<input checked="" type="checkbox"/>	11 0112F-02_F74_2_(MoGene-1_0-st-v1).CEL				x			11	2012-02-16T17:35:40Z
<input type="checkbox"/>	12 0112F-02_V1_(MoGene-1_0-st-v1).CEL							12	2012-02-15T21:49:46Z
<input type="checkbox"/>	13 0112F-02_V3_(MoGene-1_0-st-v1).CEL							13	2012-02-15T20:21:25Z
<input type="checkbox"/>	14 0112F-02_V4_(MoGene-1_0-st-v1).CEL							14	2012-02-15T21:41:15Z
<input type="checkbox"/>	15 0112F-02_V5_(MoGene-1_0-st-v1).CEL							15	2012-02-15T20:30:15Z
<input type="checkbox"/>	16 0112F-02_V6_(MoGene-1_0-st-v1).CEL							16	2012-02-15T20:48:07Z
<input type="checkbox"/>	17 0112F-02_V7_(MoGene-1_0-st-v1).CEL							17	2012-02-15T21:58:35Z

The columns named \*1, \*2, ... indicate the calls from the different outlier detection methods:

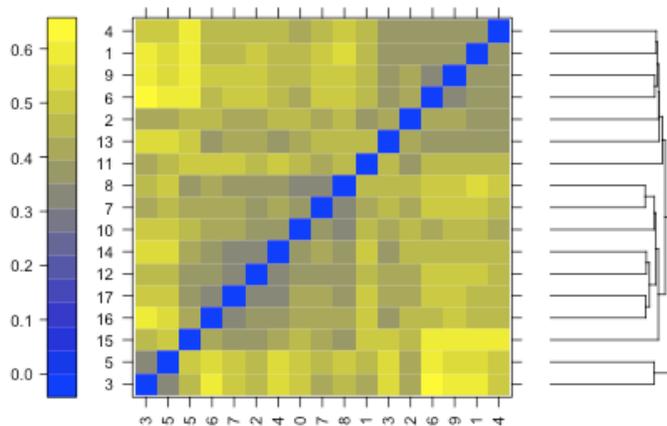
1. outlier detection by [Distances between arrays](#)
2. outlier detection by [Boxplots](#)
3. outlier detection by [Relative Log Expression \(RLE\)](#)
4. outlier detection by [Normalized Unscaled Standard Error \(NUSE\)](#)
5. outlier detection by [MA plots](#)
6. outlier detection by [Spatial distribution of M](#)

The outlier detection criteria are explained below in the respective sections. Arrays that were called outliers by at least one criterion are marked by checkbox selection in this table, and are indicated by highlighted lines or points in some of the plots below. By clicking the checkboxes in the table, or on the corresponding points/lines in the plots, you can modify the selection. To reset the selection, reload the HTML page in your browser.

At the scope covered by this software, outlier detection is a poorly defined question, and there is no 'right' or 'wrong' answer. These are hints which are intended to be followed up manually. If you want to automate outlier detection, you need to limit the scope to a particular platform and experimental design, and then choose and calibrate the metrics used.

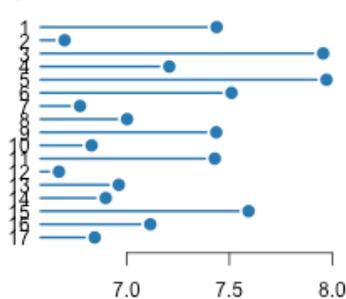
## Section 1: Between array comparison

### - Figure 1: Distances between arrays.



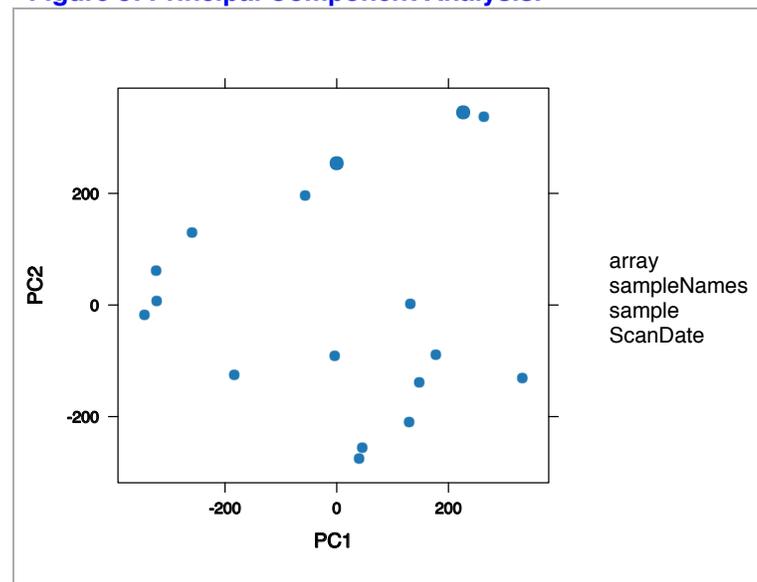
**Figure 1** ([PDF file](#)) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance  $d_{ab}$  between two arrays  $a$  and  $b$  is computed as the mean absolute difference (L<sub>1</sub>-distance) between the data of the arrays (using the data from all probes without filtering). In formula,  $d_{ab} = \text{mean } |M_{ai} - M_{bi}|$ , where  $M_{ai}$  is the value of the  $i$ -th probe on the  $a$ -th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays,  $S_a = \sum_b d_{ab}$  was exceptionally large. No such arrays were detected.

### - Figure 2: Outlier detection for Distances between arrays.



**Figure 2** ([PDF file](#)) shows a bar chart of the sum of distances to other arrays  $S_a$ , the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 8.33 was determined, which is indicated by the vertical line. None of the arrays exceeded the threshold and was considered an outlier.

### - Figure 3: Principal Component Analysis.



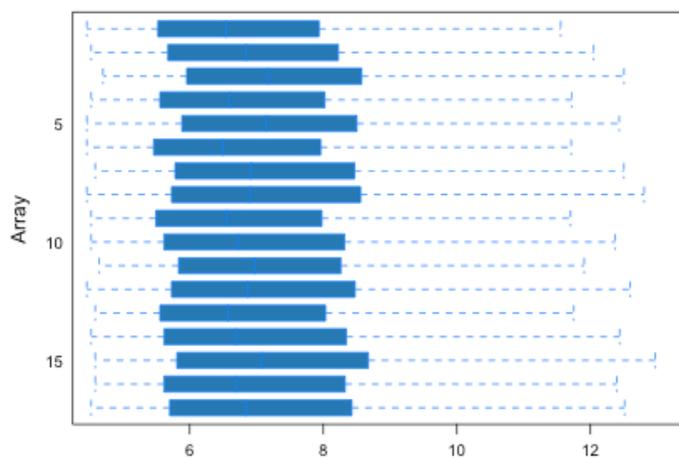
**Figure 3** ([PDF file](#)) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor (you can indicate such a factor by color using the 'intgroup' argument), or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.

Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of

each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

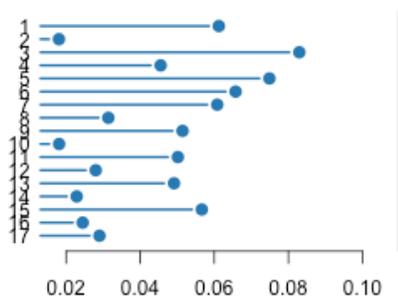
## Section 2: Array intensity distributions

### - Figure 4: Boxplots.



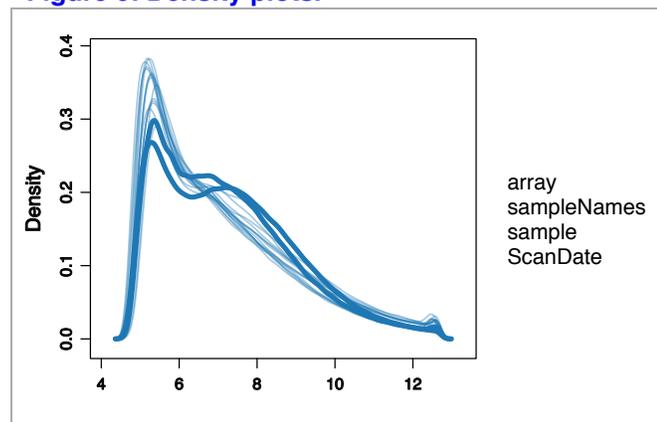
**Figure 4** ([PDF file](#)) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic  $K_a$  between each array's distribution and the distribution of the pooled data.

### - Figure 5: Outlier detection for Boxplots.



**Figure 5** ([PDF file](#)) shows a bar chart of the Kolmogorov-Smirnov statistic  $K_a$ , the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 0.11 was determined, which is indicated by the vertical line. None of the arrays exceeded the threshold and was considered an outlier.

### - Figure 6: Density plots.

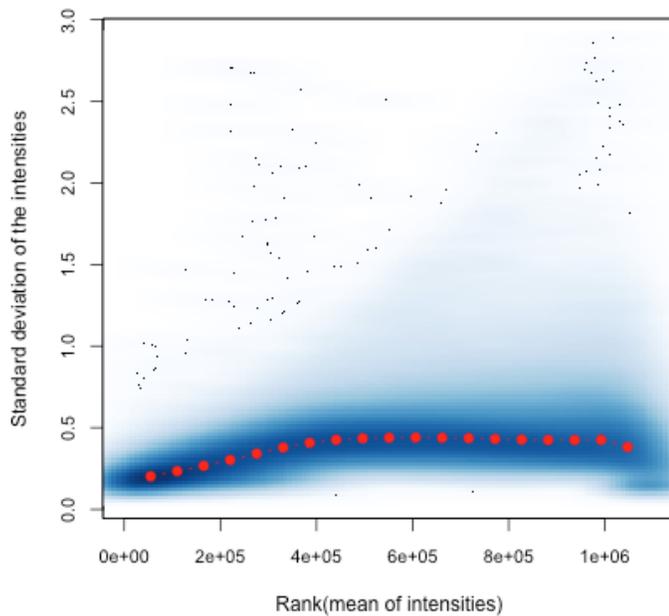


**Figure 6** ([PDF file](#)) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's

distribution to the right. Lack of signal diminishes its right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

### Section 3: Variance mean dependence

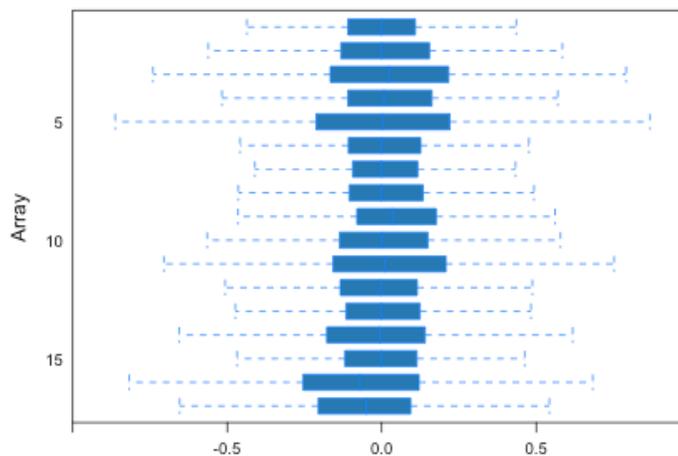
- **Figure 7: Standard deviation versus rank of the mean.**



**Figure 7** ([PDF file](#)) shows a density plot of the standard deviation of the intensities across arrays on the y-axis versus the rank of their mean on the x-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

### Section 4: Affymetrix specific plots

- **Figure 8: Relative Log Expression (RLE).**



**Figure 8** ([PDF file](#)) shows the *Relative Log Expression (RLE)* plot. Arrays whose boxes are centered away from 0 and/or are more spread out are potentially problematic. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic  $R_a$  between each array's RLE values and the pooled, overall distribution of RLE values.

- **Figure 9: Outlier detection for Relative Log Expression (RLE).**

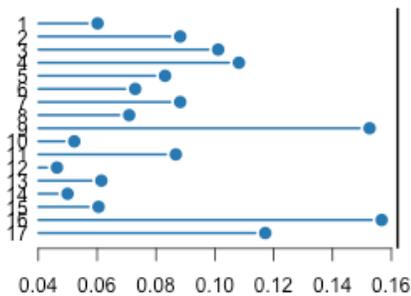


Figure 9 (PDF file) shows a bar chart of the Kolmogorov-Smirnov statistic  $R_a$  of the RLE values, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 0.162 was determined, which is indicated by the vertical line. None of the arrays exceeded the threshold and was considered an outlier.

- Figure 10: Normalized Unscaled Standard Error (NUSE).

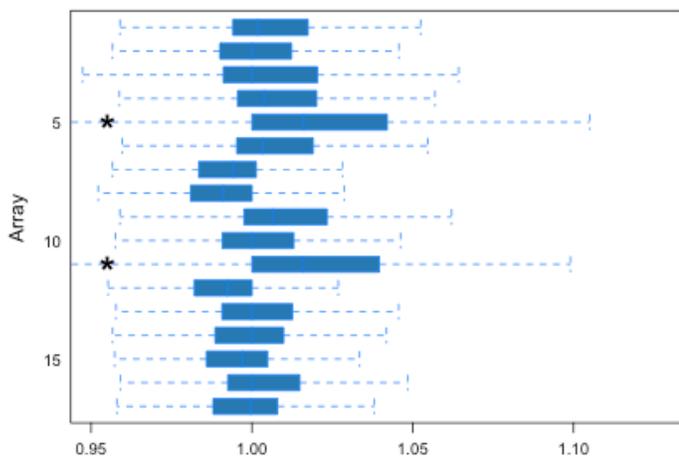


Figure 10 (PDF file) shows the Normalized Unscaled Standard Error (NUSE) plot. For each array, the boxes should be centered around 1. An array where the values are elevated relative to the other arrays is typically of lower quality. Outlier detection was performed by computing the 75% quantile  $N_a$  of each array's NUSE values and looking for arrays with large  $N_a$ .

- Figure 11: Outlier detection for Normalized Unscaled Standard Error (NUSE).

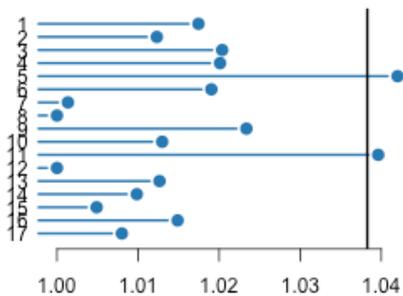
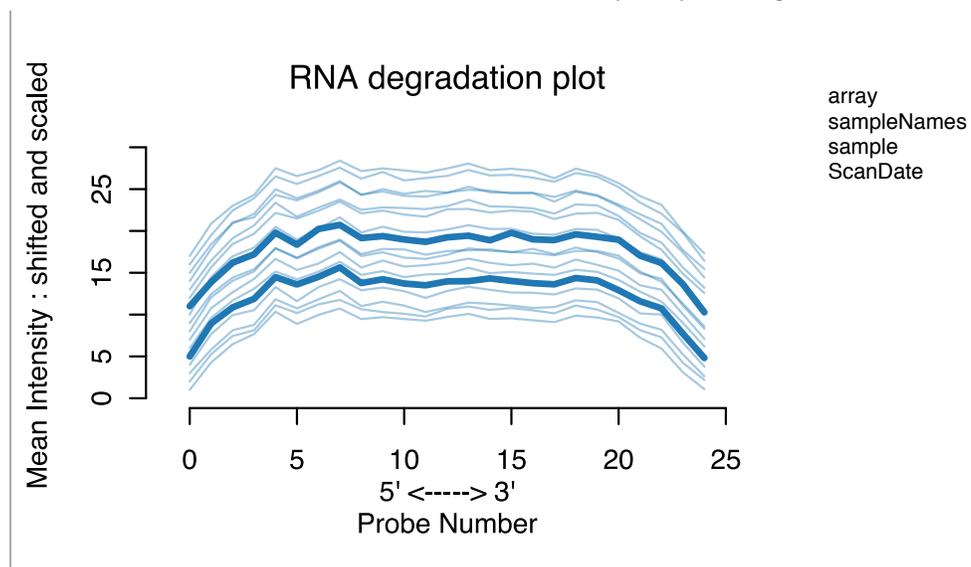


Figure 11 (PDF file) shows a bar chart of the  $N_a$ , the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 1.04 was determined, which is indicated by the vertical line. 2 arrays exceeded the threshold and were considered outliers.

- Figure 12: RNA digestion plot.

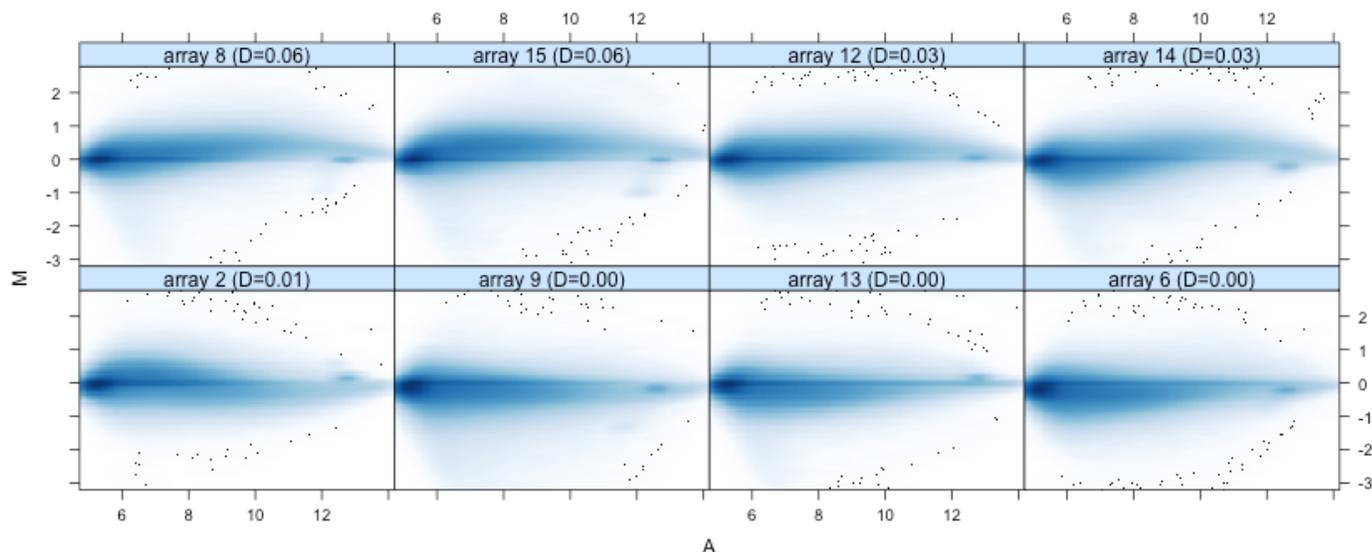




**Figure 12** ([PDF file](#)) shows the *RNA digestion* plot. The shown values are computed from the preprocessed data (after background correction and quantile normalisation). Each array is represented by a single line; move the mouse over the lines to see their corresponding sample names. The plot can be used to identify array(s) that have a slope very different from the others. This could indicate that the RNA used for that array has been handled differently from what was done for the other arrays.

## Section 5: Individual array quality

### - Figure 13: MA plots.



**Figure 13** ([PDF file](#)) shows MA plots. M and A are defined as:

$$M = \log_2(I_1) - \log_2(I_2)$$

$$A = 1/2 (\log_2(I_1) + \log_2(I_2)),$$

where  $I_1$  is the intensity of the array studied, and  $I_2$  is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the  $M = 0$  axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).

Outlier detection was performed by computing Hoeffding's statistic  $D_a$  on the joint distribution of A and M for each array. Shown are first the 4 arrays with the highest values of  $D_a$ , then the 4 arrays with the lowest values. The value of  $D_a$  is shown in the panel headings. 0 arrays had  $D_a > 0.15$  and were marked as outliers. For more information on Hoeffding's  $D$ -statistic, please see the manual page of the function `hoef.fd` in the `Hmisc` package.

### - Figure 14: Outlier detection for MA plots.

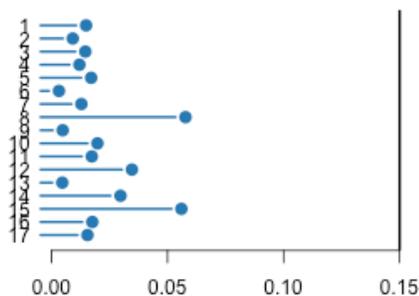
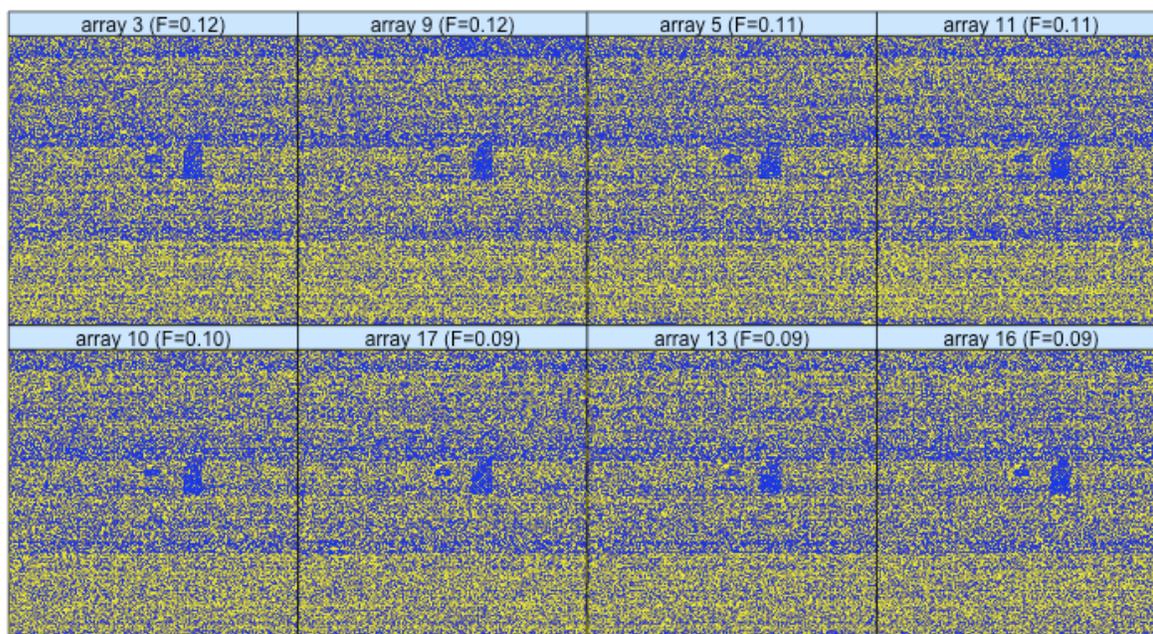


Figure 14 (PDF file) shows a bar chart of the  $D_a$ , the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. A threshold of 0.15 was used, which is indicated by the vertical line. None of the arrays exceeded the threshold and was considered an outlier.

- Figure 15: Spatial distribution of M.



M

Figure 15 (PDF file) shows false color representations of the arrays' spatial distributions of feature intensities ( $M$ ). Normally, when the features are distributed randomly on the arrays, one expects to see a uniform distribution; control features with particularly high or low intensities may stand out. The color scale is proportional to the ranks of the probe intensities. Note that the rank scale has the potential to amplify patterns that are small in amplitude but systematic within an array. It is possible to switch off the rank scaling by modifying the `argumentscale` in the call of the `aqm.spatial` function.

Outlier detection was performed by computing  $F_a$ , the sum of the absolute value of low frequency Fourier coefficients, as a measure of large scale spatial structures. Shown are first the 4 arrays with the highest values of  $F_a$ , then the 4 arrays with the lowest values. The value of  $F_a$  is shown in the panel headings.

- Figure 16: Outlier detection for Spatial distribution of M.

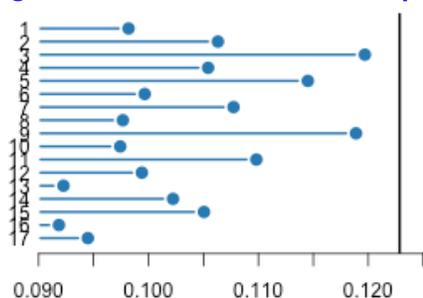


Figure 16 (PDF file) shows a bar chart of the  $F_a$ , the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 0.123 was determined, which is indicated by the

vertical line. None of the arrays exceeded the threshold and was considered an outlier.

---

This report has been created with arrayQualityMetrics 3.24.0 under R version 3.2.0 (2015-04-16).

---

(Page generated on Thu Jul 16 11:01:37 2015 by [hwriter](#) )