

# arrayQualityMetrics report for salinity\_ExpressionSet\_unfiltered

- [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: Individual array quality](#)
  - MA plots

## Browser compatibility

This report uses recent features of HTML 5. Functionality has been tested on these browsers: Firefox 10, Chrome 17, Safari 5.1.2

## - Array metadata and outlier detection overview

array	sampleNames	*1	*2	*3	sample_type	sample_name
<input checked="" type="checkbox"/>	1 c253295110048_ar2_jun1912_A_SL			x	sal25ppt	c253295110048_ar2_jun1912_A_SL
<input checked="" type="checkbox"/>	2 c253295110048_ar1_jun1912_B_SL	x		x	sal25ppt	c253295110048_ar1_jun1912_B_SL
<input type="checkbox"/>	3 c253295110048_ar4_jun1912_C_SL				sal25ppt	c253295110048_ar4_jun1912_C_SL
<input type="checkbox"/>	4 c253295110048_ar3_jun1912_A_L				sal30ppt	c253295110048_ar3_jun1912_A_L
<input type="checkbox"/>	5 c253295110048_ar6_jun1912_B_L				sal30ppt	c253295110048_ar6_jun1912_B_L
<input type="checkbox"/>	6 c253295110048_ar5_jun1912_C_L				sal30ppt	c253295110048_ar5_jun1912_C_L
<input type="checkbox"/>	7 c253295110048_ar8_jun1912_A_C				sal35ppt	c253295110048_ar8_jun1912_A_C
<input type="checkbox"/>	8 c253295110048_ar7_jun1912_B_C				sal35ppt	c253295110048_ar7_jun1912_B_C
<input type="checkbox"/>	9 c253295110054_ar2_jul1112_C_C				sal35ppt	c253295110054_ar2_jul1112_C_C

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

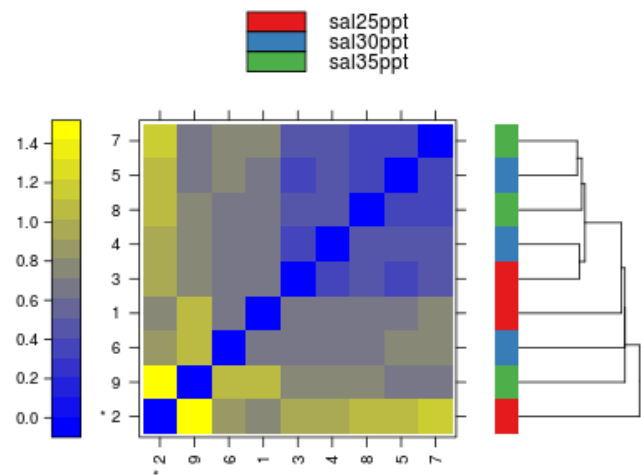
1. outlier detection by [Distances between arrays](#)
2. outlier detection by [Boxplots](#)
3. outlier detection by [MA plots](#)

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.

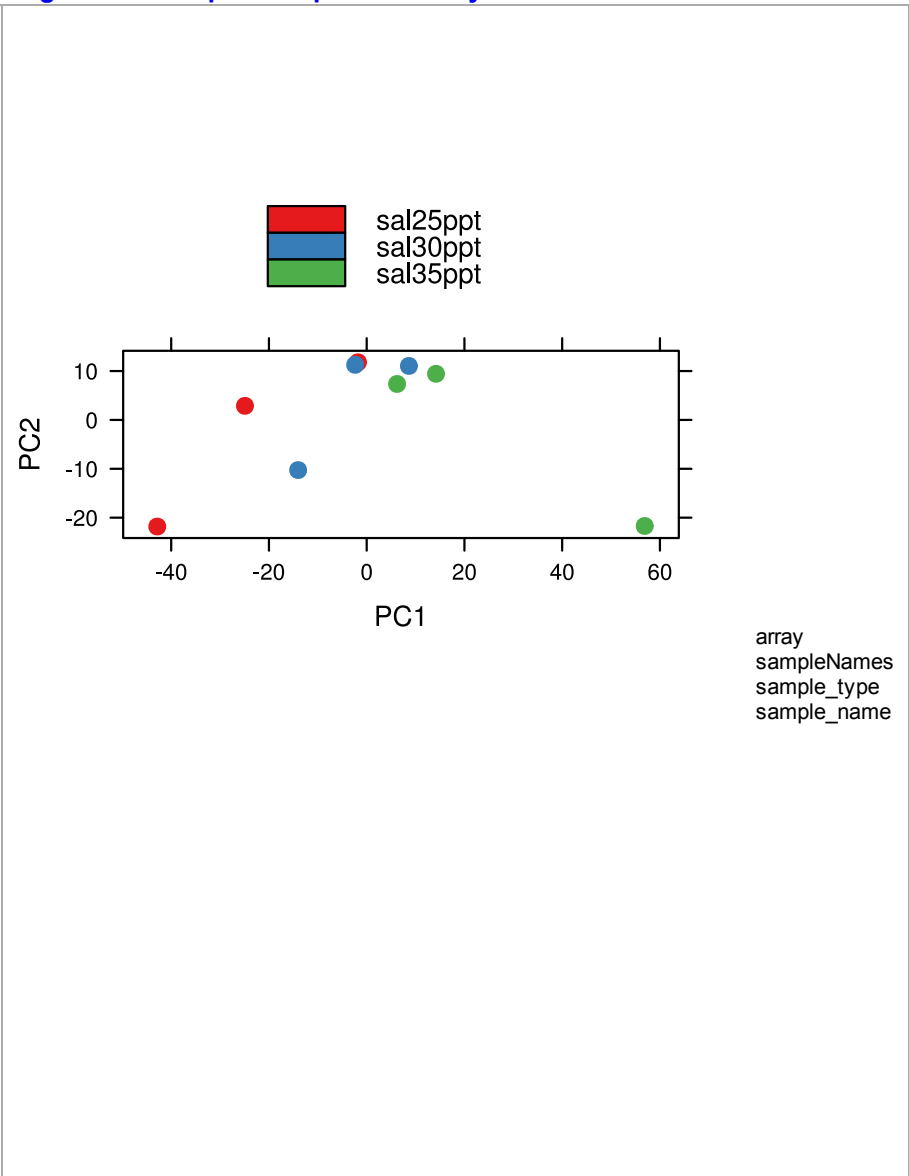
## Section 1: Between array comparison

### - 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 ( $L_1$ -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. One such array was detected, and it is marked by an asterisk, \*.

+ **Figure 2: Outlier detection for Distances between arrays.**  
- **Figure 3: Principal Component Analysis.**



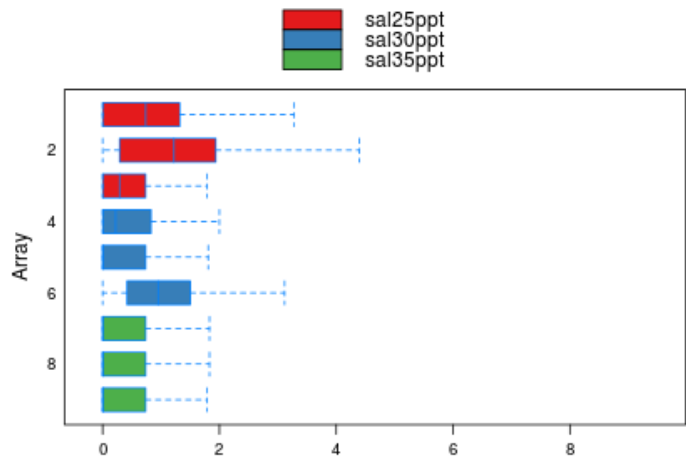


**Figure 3** ([PDF file](#)) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor, or according to unintended causes such as batch effects. Move the mouse over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

Note: the figure is static - enhancement with interactive effects failed. This is either due to a version incompatibility of the 'SVGAnnotation' R package and your version of 'Cairo' or 'libcairo', or due to plot misformatting. Please consult the Bioconductor mailing list, or contact the maintainer of 'arrayQualityMetrics' with a reproducible example in order to fix this problem.

Section 2: Array intensity distributions

- **Figure 4: Boxplots.**

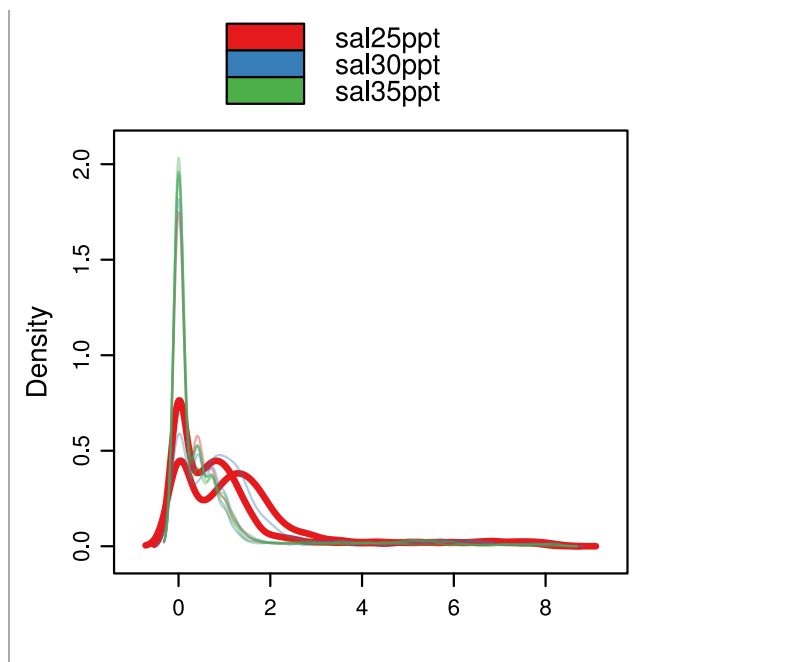


**Figure 4** ([PDF file](#)) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may 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 6: Density plots.**



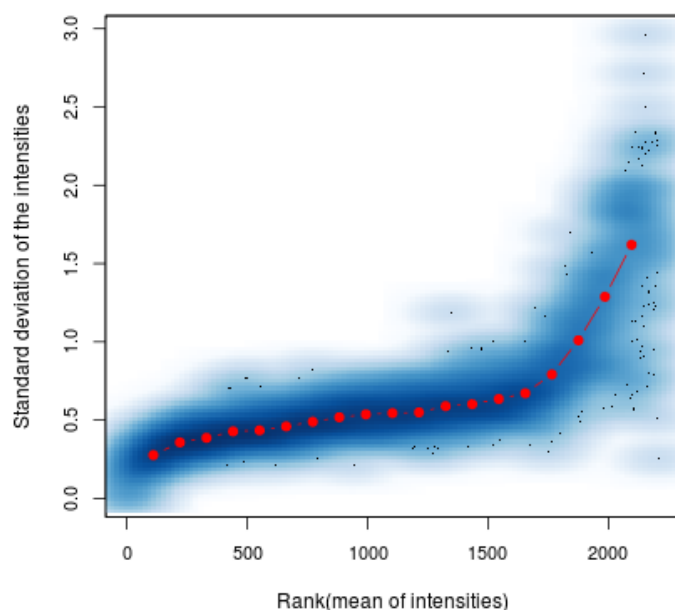


**Figure 6** ([PDF file](#)) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions 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.

Note: the figure is static - enhancement with interactive effects failed. This is either due to a version incompatibility of the 'SVGAnnotation' R package and your version of 'Cairo' or 'libcairo', or due to plot misformatting. Please consult the Bioconductor mailing list, or contact the maintainer of 'arrayQualityMetrics' with a reproducible example in order to fix this problem.

### Section 3: Variance mean dependence

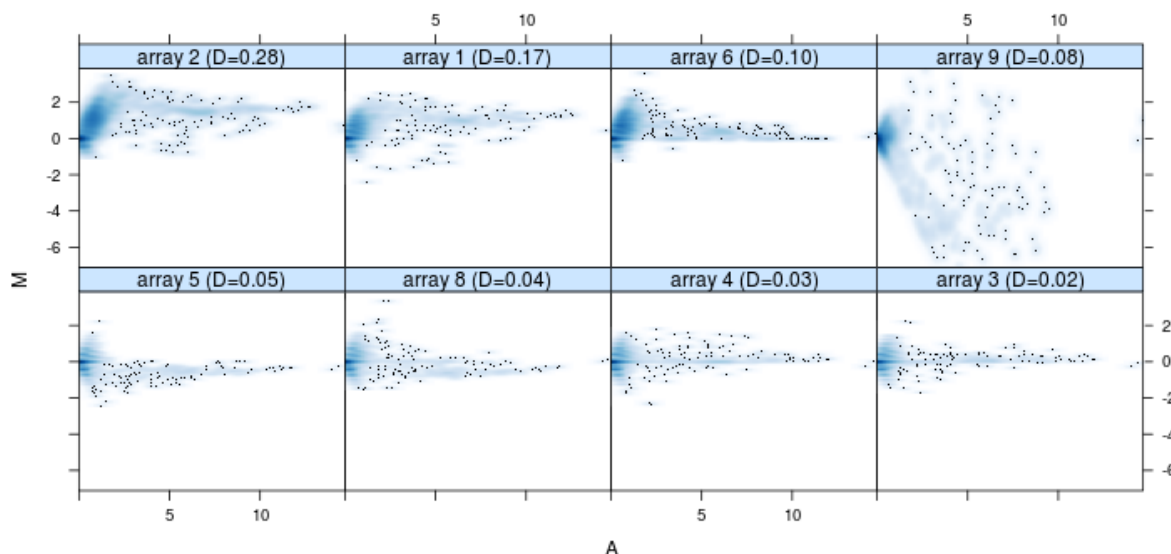
- **Figure 7: Standard deviation versus rank of the mean.**



**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: Individual array quality

### - Figure 8: MA plots.



**Figure 8** ([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. 2 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 `hoeffd` in the `Hmisc` package.

### + Figure 9: Outlier detection for MA plots.

This report has been created with arrayQualityMetrics 3.18.0 under R version 3.0.2 (2013-09-25).

(Page generated on Tue Apr 7 21:22:51 2015 by [hwriter](#) )