

# Validation of the Microarray Data Analyzer Software MADA using reference data sets

<b>1</b>	<b>FUNCTIONALITY OF MADA</b>	<b>2</b>
1.1	Validation based on a data set taken from Peplies et al., 2004 (Identification)	2
1.2	Validation of the MADA outlier test	4
1.3	Validation based on a data set taken from Helmann et al., 2003 (Gene expression profiling)	6
<b>2</b>	<b>COMPARISON TO SIMILAR SOFTWARE TOOLS</b>	<b>9</b>
2.1	<b>MarC-V (Schageman et al., 2002)</b>	<b>9</b>
2.1.1	Based on the data set from Peplies et al., 2004	9
2.1.2	Based on the data set from Helmann et al., 2003	9
2.2	<b>BRB-ArrayTools</b>	<b>12</b>
2.2.1	Based on the data set from Peplies et al., 2004	12
2.2.2	Based on the data set from Helmann et al., 2003	12
<b>3</b>	<b>CONCLUSIONS</b>	<b>15</b>
<b>4</b>	<b>REFERENCES</b>	<b>16</b>

---

## UPDATE

02.2007 - Additional information according the calculation speed of MADA 2.0 was added at page 3, 7 + 15 and is highlighted by red color.

---

Andreas Ellrott

Max Planck Institute for Marine Microbiology  
Microbial Genomics Group  
Celsiusstr. 1, D-28359 Bremen, Germany

<http://www.megx.net/mada>

08.2005 / 02.2007

## 1 Functionality of MADA

### 1.1 Validation based on a data set taken from Peplies et al., 2004 (Identification)

In the study of Peplies et al. 2004, a DNA microarray for the detection of marine bacteria based on 16S rRNA-targeted probes was set up and validated. Figure 1 shows the image of the scanned microarray that corresponds to the results shown in Fig. 1B of the publication. The raw data file can be downloaded at <http://www.megx.net/mada>.

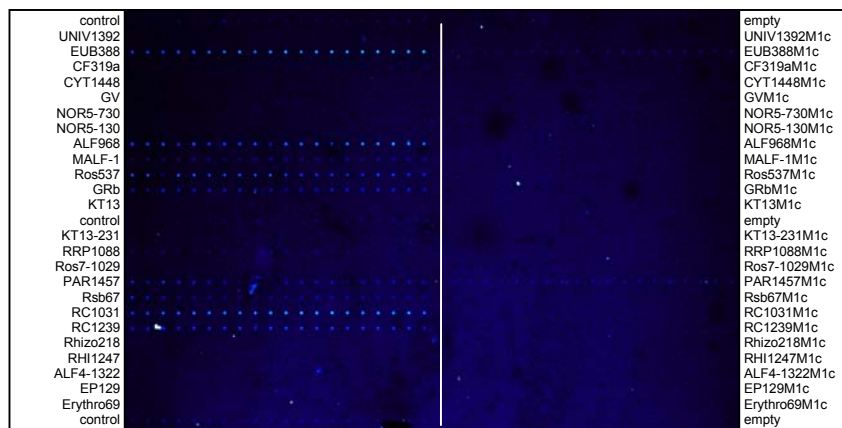


Figure 1. Image of the scanned microarray that correspond to the experiment shown in Fig. 1B of the publication of Peplies et al., 2004.

The raw data file was imported into MADA. Channel 1 and 2 were extracted using the QuantArray extraction filter. The calculation settings were as followed: local background correction, signal significance test ( $t\text{-score} = 2$ ), no outlier test (advanced options), mean of replicates from at least ten valid signals (twenty replicates in total for each probe), no normalization, StDev from replicates, and ratio & intensity disabled (advanced options).

Figure 2 shows the spot signals which were assigned as significantly higher than the local background. No significant signals were found for the mismatch controls (located right to the vertical line) as also reported by Peplies et al. who processed the raw data manually.

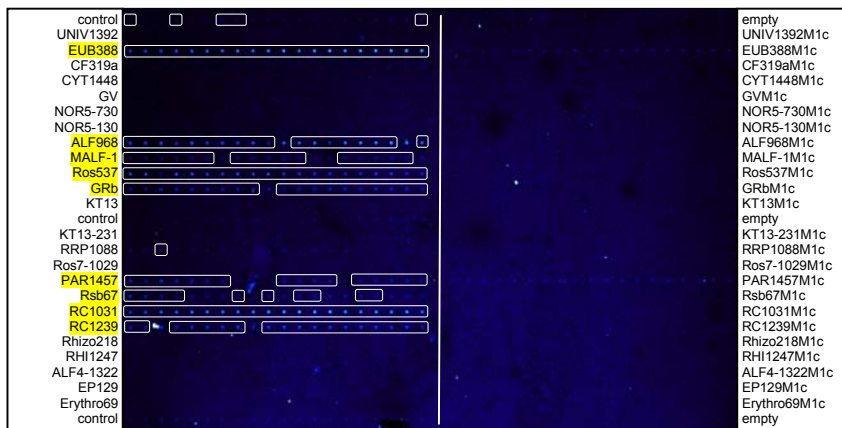


Figure 2. Spot signals that were assigned as significantly above the local background by MADA (framed) with a probability for false-positive results of < 2,5%. Probes are highlighted for which the mean of replicates was calculated according to the setting 'at least ten valid signals' within the twenty replicates.

Figure 3 was created from the processed raw data set using the MADA 'Plot Chart', 'Mean of replicates' function for selected values of Channel 2. It represents an exact match to the corresponding figure published by Peplies et al.

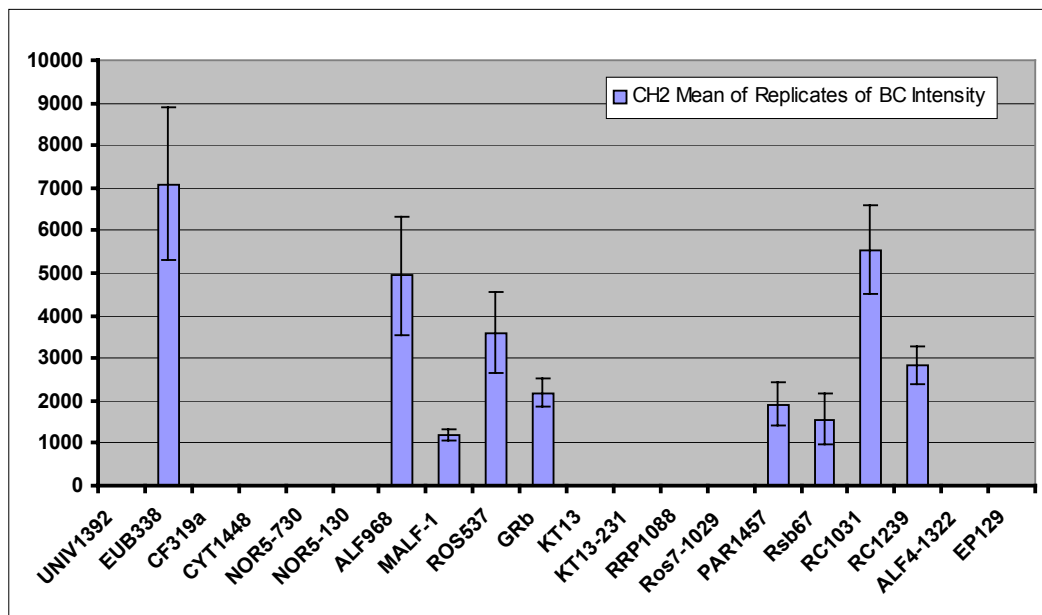


Figure 3. MADA 'Mean of replicate' plot for the raw data set taken from Peplies et al, 2004.

The analysis was done with a Pentium III 700MHz and 512MB of RAM using Excel 97 with the operating system WinNT. The total computing time was ~1min for import, extraction, calculation and visualization.

Update 02.2007: By using MADA 2.0 on the same computer system, the total computing time for import, extraction, calculation and visualization was ~0.66min.

## 1.2 Validation of the MADA outlier test

A modified version of the original raw data file (compare section 1.1) was created by manual insertion of five outlying data points within selected probe replicates. For probe MALF-1, the signal intensity of one replicate was increased 5-fold to simulate a hybridization artifact such as dirt on the slide surface. Single replicates with 3-fold and 2-fold reduced signal intensity were introduced for probes RC1031 and RC1239, respectively, to simulate spotting artifacts. For probe GRb, the signal intensity of two replicates was 4-fold increased and 2-fold decreased, respectively. The modified raw data file can be found at <http://www.megx.net/mada>.

Figure 4A was calculated without using the MADA outlier test to demonstrate the effect of the introduced outliers on the data quality (indicated by the high standard deviations for the corresponding probes). Subsequently, the data set was processed using the MADA outlier test with a setting for the deviation difference of  $>30\%$  (Fig. 4B). All outlying data points have been identified by MADA.

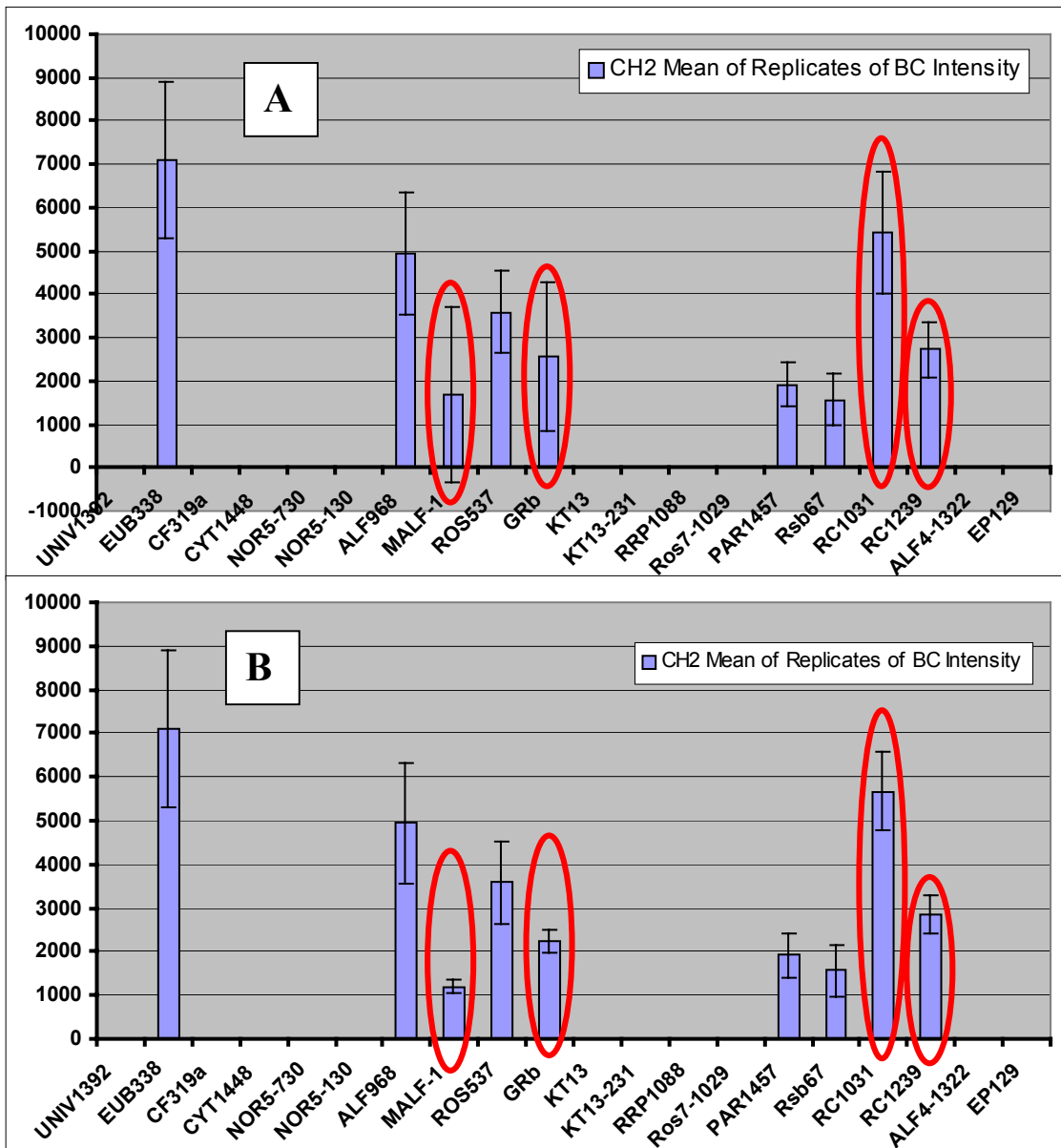


Figure 4. Signals and the corresponding standard deviations calculated by MADA from the modified raw data file containing outlying data points within selected probe replicates. (A) Without performing the MADA outlier test. (B) Using the MADA outlier test with a setting for the deviation difference of >30%.

### 1.3 Validation based on a data set taken from Helmann et al., 2003 (Gene expression profiling)

In the study of Helmann et al., 2003 the expression level of approx. 3700 *Bacillus subtilis* genes was compared for a *perR* mutant and a wild type strain by microarray hybridization of six replicated arrays each containing 4608 spots. A number of ~75 genes was identified as significantly upregulated (3-fold) in the *perR* mutant (Table 1 of the corresponding publication). The raw data files (Bacillus PerR set, ExptID 25863 – 25868) are available at the Stanford Microarray Database (SMD) at [http://smd.stanford.edu/cgi-bin//publication/viewPublication.pl?pub\\_no=205](http://smd.stanford.edu/cgi-bin//publication/viewPublication.pl?pub_no=205).

Since the data organization in the raw data files does not fit to any common structure, a specific extraction filter was created using MADA's integrated 'Edit filter'-tool (Fig. 5). The filter can be found at <http://www.megx.net/mada>.

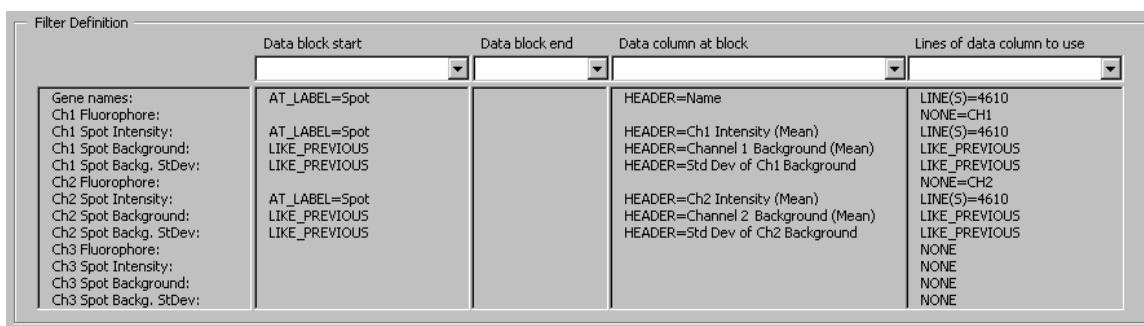


Figure 5. Settings for the MADA extraction filter created for the extraction of the raw data file provided by Helmann et al., 2003.

The six raw data files (25863.xls to 25868.xls) were imported into MADA as 'Raw Data 1' to 'Raw Data 6'. Data extraction was done for channel 1 and 2 using the extraction filter shown in Fig. 5 and the option 'consecutive grouping of replicates via names'. Calculation was done with the following settings: local background correction, signal significance test with a probability for false-positive results of <5% (t-score = 1.6), no outlier test (advanced options), mean of replicates from at least one valid signal, normalization at own channels according to the 'mean of overall signal intensity' option, StDev from replicates and ratio & intensity.

Because the channels of the two fluorescent dyes were switched for the original raw data files 25863 to 65 and 25866 to 68, the channel assignment for ratio & intensity in MADA was also switched (Ch2/Ch1 for the first three and Ch1/Ch2 for the last three hybridizations).

MADA was not able to combine the forty replicates of the probe called 'GENOME' to a single value since the corresponding Excel formula for 'mean of replicates' exceeds the Excel limitation of the maximum formula length.

For visualization of the results, the MADA RI-Plot was used with only the upregulated genes selected according to a regulation factor of 3.

In Table 1 the expression ratios calculated by MADA are shown in comparison to the results of Hellmann et al., 2003 (Table 1 of the publication). The ratios calculated by MADA correspond to the foldinduction values of the genes given by Hellmann et al. with an accuracy of 97.4%  $\pm$ 2.4%. The divergence is presumably due to slight differences in the calculation and normalization procedures which were not specified in detail by Hellmann et al. Three genes are not included in the comparison, the *zosA* gene was missing in the raw data files and for the *yybI* and *ytpQ* gene no differential expression according to the 3-fold cutoff could be detected.

With a Pentium III 700MHz and 512MB of RAM using Excel 97 with the operating system WinNT the computing time for each raw data file including import, extraction and calculation was  $\sim$  4.5min. A Pentium IV with 1GB of RAM, using Excel 2002 (Office XP) at WinXP was able to handle this task in  $\sim$  2 min.

Update 02.2007: MADA at version 2.0 was able to handle this task in  $\sim$ 0.33min on the Pentium III 700MHz, 512MB Ram using Excel 97 at Windows NT and in  $\sim$ 0.15min on the Pentium IV 3GHz with 1GB of RAM, using Excel 2002 at Windows XP.

Helmann et al. 2003, Tab.1		MADA						Ratio values transformed to foldinduction						Average	Accuracy [%]
Gene(s)	Foldinduction	Ratio values logarithm base 2						Ratio values transformed to foldinduction							
		Ch2 / Ch1			Ch1 / Ch2			Ch2 / Ch1			Ch1 / Ch2				
		Experiment 25663	Experiment 25664	Experiment 25665	Experiment 25666	Experiment 25667	Experiment 25668	Experiment 25663	Experiment 25664	Experiment 25665	Experiment 25666	Experiment 25667	Experiment 25668		
comER	9.9	3.29	3.37	2.99	3.10		9.76	10.32	7.94	8.55		9.14	92.3		
yybF	9.81	3.38	3.24	3.31	3.36	3.21	3.31	10.42	9.43	9.91	10.26	9.28	9.94	9.87	99.4
mrgA	9.64	3.26	3.25	3.23	3.24	3.34	3.35	9.58	9.53	9.41	9.42	10.14	10.18	9.71	99.3
ywfM	8.94	3.44	3.37		3.04	2.92		10.85	10.36		8.21	7.59		9.25	96.6
comEA	8.27	2.94	2.93	2.86	3.25		3.28	7.67	7.60	7.25	9.52		9.73	8.35	99.0
comGA		3.14	3.05	3.71	3.51	3.42		8.81	8.29	13.10	11.39	10.73		10.46	
comGB		3.15	2.86	2.90	3.19	3.20	3.87	8.87	7.27	7.46	9.11	9.16	14.60	9.41	
comGC		2.74	2.71	3.41	3.15	3.22		6.66	6.54	10.66	8.87	9.31		8.41	
comGD		3.15	3.17	3.03	3.12	3.15	3.94	8.87	8.97	8.19	8.72	8.86	15.36	9.83	
comGE		2.86	2.84	2.79	3.44			7.27	7.18	6.92	10.86			8.06	
comGF		2.82	2.91	2.75		2.89	3.54	7.05	7.53	6.72	1.00	7.40	11.66	6.89	
comGG		1.95	1.92	1.91	2.39	2.27	2.30	3.86	3.78	3.75	5.25	4.82	4.91	4.39	
comG(ABCDEF)	7.93 ± 1.74													8.21 ± 2.06	96.6
melA	7.31	2.87	2.80	2.88	3.23	3.03	3.10	7.32	6.98	7.36	9.41	8.14	8.56	7.96	91.8
cwJ	7.64	3.01	3.04	2.84	2.88			8.04	8.22	7.14	7.38			7.70	99.3
gbsA		3.04		2.97	2.96	2.91	3.37	8.21		7.82	7.80	7.49	10.31	8.32	
gbsB			2.53	2.51	2.56	2.80			5.77	5.69	5.92	6.95		6.08	
gbsAB	7.62 ± 0.52													7.20 ± 1.59	94.5
msmR		2.23	2.20	2.26	2.49	2.38	2.33	4.70	4.60	4.78	5.62	5.20	5.03	4.99	
msmE		3.24	3.21	3.30	2.88	2.90	2.89	9.46	9.23	9.83	7.37	7.45	7.40	8.46	
amyD		2.97	2.84	2.96	3.18	3.07	2.98	7.85	7.15	7.78	9.06	8.39	7.87	8.02	
amyC		3.26	3.15	3.27	2.91	2.99	2.92	9.57	8.87	9.62	7.50	7.95	7.59	8.52	
msmRE amyDC	7.44 ± 1.67													7.50 ± 1.69	99.3
dppA		2.48	2.52	2.48	2.76	2.73	3.17	5.58	5.72	5.58	6.79	6.65	9.01	6.55	
dppB		2.84		2.71		3.25		7.14		6.54		9.51		7.73	
dppC		2.68	2.48	2.65	3.13	2.99		6.41	5.60	6.29	8.73	7.95		7.00	
dppD		2.54	2.47	2.35	3.13	3.12	3.16	5.81	5.53	5.10	8.78	8.71	8.91	7.14	
dppE		2.72	2.62	2.76	2.71	2.63	3.11	6.59	6.14	6.76	6.52	6.19	8.64	6.81	
dppABCDE	7.04 ± 0.81													7.05 ± 0.44	99.9
zosA	5.97	not available in the raw data sets													
appD		2.56	2.48	2.60	2.53	2.47	2.73	5.91	5.58	6.06	5.77	5.53	6.64	5.91	
appF		2.39	2.21	2.33	2.58	2.54	2.49	5.24	4.62	5.03	6.00	5.81	5.61	5.39	
appDF	5.53 ± 0.34													5.65 ± 0.37	97.9
katA	5.34	2.45	2.40	2.43	2.34	2.37	2.52	5.45	5.28	5.40	5.06	5.18	5.72	5.35	99.9
livB		2.96	2.93	2.88	2.70	2.77	3.03	7.79	7.63	7.36	6.49	6.81	8.19	7.38	
livC			2.16		1.95	1.94	2.16		4.46		3.87	3.85	4.47	4.16	
leuB			2.03	2.00	1.97				4.09	4.01	3.90			4.00	
leuC		1.98	2.02	2.11	2.39	2.33	2.33	3.96	4.05	4.33	5.25	5.02	5.02	4.61	
livBC leuBC	5.01 ± 0.71													5.04 ± 1.58	99.5
ureA			2.44	2.29		2.55			5.42	4.89		5.84		5.39	
ureB			2.19	2.19		2.24			4.56	4.56		4.73		4.61	
ureC		1.81		1.71	1.96	1.92		3.50		3.27	3.90	3.79		3.61	
ureABC	4.96 ± 1.4													4.54 ± 0.89	91.5
yckD			1.98		2.10	1.96			3.95		4.28	3.88		4.03	
yckE		2.33	2.13	2.24	2.57	2.45	2.40	5.02	4.38	4.72	5.92	5.47	5.29	5.13	
yckDE	4.96 ± 0.1													4.58 ± 0.78	92.4
yybI	4.92			0.39		0.17				1.31		1.12		not expressed	
appB		2.01	1.81	1.91	2.33	2.28	2.20	4.04	3.51	3.76	5.04	4.84	4.58	4.30	
appC		2.38	2.42	2.30	2.32	2.45	2.93	5.20	5.34	4.92	4.98	5.47	7.64	5.59	
appBC	4.98 ± 0.9													4.94 ± 0.92	99.3
yxbC	4.6	2.38	2.17	2.24	2.10	2.04	2.51	5.22	4.51	4.73	4.28	4.13	5.71	4.76	96.6
yxbB		2.21	2.05	2.04	2.37	2.15	2.29	4.61	4.14	4.11	5.18	4.44	4.88	4.56	
yxbA		2.15	2.03	2.19	1.92	2.06		4.42	4.09	4.55	3.79	4.17		4.20	
yxbBA	4.5 ± 0.1													4.38 ± 0.25	97.4
yufN	4.44			2.24	2.19					4.72	4.56			4.64	95.7
yvaW		1.62	1.53	1.60	1.79	1.72	1.05	3.08	2.89	3.04	3.45	3.30	2.07	2.97	
yvaY		2.50	2.36	2.50	2.70	2.60	2.58	5.67	5.14	5.65	6.48	6.06	5.99	5.83	
yvaWY	4.36 ± 0.2													4.40 ± 2.02	99.0
yhdG	4.27		1.63	1.89	2.61	2.23			3.10	3.70	6.10	4.68		4.39	97.2
slpT	4.25	2.14	1.79	2.06		2.17		4.41	3.46	4.16		4.49		4.13	97.1
ygbA	4.18	2.00	2.06	1.98	2.05	2.04	2.31	4.00	4.16	3.94	4.15	4.10	4.98	4.22	99.0
ppsA		1.84	1.61	1.69		1.98	1.95	3.58	3.05	3.22		3.94	3.87	3.53	
ppsB		1.78	1.55	1.58	2.09	1.00	2.00	3.44	2.93	2.98	4.26	2.00	3.99	3.27	
ppsC		2.06	2.03	2.02	1.99	1.66		4.17	4.09	4.07	3.97	3.16		3.89	
ppsD		2.36		2.18	1.67	1.73		5.13		4.52	3.19	3.31		4.04	
ppsE		2.05	2.09	2.03	1.93	1.90	2.01	4.14	4.25	4.09	3.82	3.74	4.03	4.01	
ppsABCDE	3.93 ± 0.52													3.75 ± 0.34	95.4
yuiA	3.9			1.87		2.06				3.66		4.17		3.91	99.7
ykvK		1.63	1.74	1.54	1.82	1.69	2.22	3.10	3.33	2.91	3.52	3.23	4.65	3.46	
ykvL		2.24	2.21	2.22	2.06	2.10	2.21	4.74	4.63	4.66	4.17	4.30	4.63	4.52	
ykvM		1.82	1.83	1.78	2.00	1.89	2.03	3.53	3.56	3.44	4.01	3.70	4.08	3.72	
ykvKLM	3.83 ± 0.6													3.90 ± 0.55	98.2
yheK	3.76	1.96	1.88	1.93	1.97	1.86	1.99	3.89	3.69	3.81	3.93	3.64	3.99	3.82	98.3
yisS	3.67	1.73	1.60	1.72	2.12	2.05	2.11	3.32	3.03	3.29	4.36	4.13	4.32	3.74	98.1
ykfB		1.93	1.81	1.77	2.17	1.92		3.82	3.51	3.40	4.50	3.79		3.80	
ykfC		1.62	1.47	1.53	2.09	2.00	2.10	3.07	2.77	2.88	4.25	3.99	4.28	3.54	
ykfD		1.76	1.65	1.65	1.87	1.81	2.28	3.38	3.15	3.13	3.66	3.50	4.84	3.61	
ykfBCD	3.67 ± 0.34													3.65 ± 0.14	99.5
ywcE	3.5	1.70	1.52	1.72	2.13	1.91	1.94	3.25	2.87	3.29	4.39	3.75	3.84	3.57	98.2
yoeB	3.29	1.69	1.70	1.73	1.63	1.72	1.85	3.22	3.24	3.31	3.10	3.29	3.61	3.30	99.8
yhdC	3.27	1.75	1.69	1.55	1.81	1.81	1.75	3.37	3.22	2.93	3.52	3.51	3.36	3.32	98.5
bioA		1.71	1.69	1.68	1.79	1.82	1.78	3.28	3.22	3.20	3.46	3.53	3.44	3.35	
bioF		1.44	1.34	1.39	2.08	2.00	2.03	2.71	2.53	2.63	4.22	3.99	4.07	3.36	
bioD		1.74	1.64	1.72	1.75	1.78	1.80	3.34	3.11	3.30	3.36	3.44	3.49	3.34	
bioB		1.45	1.44	1.38	2.01	1.81	1.83	2.73	2.72	2.60	4.04	3.50	3.56	3.19	
bioAFDB	3.26 ± 0.11													3.31 ± 0.08	98.4
rapH	3.21	1.44	1.38	1.48	2.00	1.88	1.93	2.71	2.61	2.78	4.01	3.68	3.81	3.27	98.3
yybG	3.14	1.76	1.75	1.62	1.49	1.65	1.67	3.38	3.36	3.08	2.80	3.14	3.19	3.16	99.5
opuBB	3.14	1.69	1.95	1.74	1.53			3.23	3.87	3.34	2.88			3.33	94.3
ytpQ	3.1	-0.04	-0.08	-0.05	-0.01	0.03	0.03	0.98	0.95	0.96	0.99	1.02	1.02	not expressed	
glfB	3.07	1.73	1.73	1.67	1.42	1.51	1.67	3.31	3.32	3.19	2.67	2.85	3.18	3.09	99.4
ywfH	3.07		1.38	1.38	1.88	1.83	1.82		2.60	2.61	3.67	3.55	3.52	3.19	96.2
yqzE	2.88		1.24	1.46	1.56	1.53	2.14		2.36	2.75	2.94	2.88	4.41	3.07	93.9
Average															97.4 ± 2.4

Table 1. Comparison of the original data published by Helmann et al., 2003 (Tab. 1 of

## **2 Comparison to similar software tools**

### **2.1 MarC-V (Schageman et al., 2002)**

MarC-V represents the most similar software tool compared to MADA since it is implemented as a VBA macro in Excel and is equipped with a graphical user interface. MarC-V was developed for normalization and visualization of single, two-color microarray experiments in the field of gene expression profiling using GenePix® data files. Data from other sources must be copied and manually pasted into the program by using a special custom version of MarC-V. Various options and visualizations can be accessed within ten Excel worksheets. The program was published by Schageman et al. in 2002 and can be downloaded at [http://pga.swmed.edu/Information/marcV\\_Info.htm](http://pga.swmed.edu/Information/marcV_Info.htm).

#### **2.1.1 Based on the data set from Peplies et al., 2004**

Since MarC-V was exclusively designed for the analysis of two-color hybridization experiments in expression profiling, the program was not able to process the data from the study of Peplies et al., 2004 in a reasonable way. Here, only the data from a single channel (Ch2 - the data from Ch1 just represent control spots) are of interest for organism identification purposes.

#### **2.1.2 Based on the data set from Helmann et al., 2003**

Since only files in the GenePix file format can be imported into the current version 2.95 of MarC-V, initially the raw data provided by Helmann et al., 2003, could not be processed by the program. Therefore, the 2.91.custom version of MarC-V was used which is provided for manual data import using the Windows 'copy and paste' function.

Before extraction of raw data columns required by MarC-V, the original raw data file had to be adapted in different ways. All spots named '6xSSC' were renamed to 'BLANK' since the program requires 'blank' spots for threshold calculations. Moreover, MarC-V requires information on spot quality, often provided by manual 'flagging' of the spots during image analysis to differentiate between 'good' and 'bad' spots. Since this information is not available in the raw data files of Helmann et al., an additional column called 'Flag' was inserted by hand with a setting of '100' for each spot which corresponds to 'good'. Subsequently, the columns 'Name', 'Flag', 'Ch1 Intensity (Mean), Channel 1 Background (Mean), 'Ch2 Intensity (Mean) and Channel 2 Background (Mean) were manually copied into MarC-V with the Excel automatic calculation deactivated.

After reactivation of the automatic calculation data processing was started. The calculation is based on local background correction, threshold adjustment, Cy3 and Cy5 ratios, and normalization according to mean log ratio by default and can not be modified by the user. Results are visualized by different plots

such as scatter plots, commonly used in gene expression analysis. Probe replicates are processed as single genes and can not be combined into a single mean or median value. For analysis of each of the six data sets the program was restarted since MarC-V includes no batch processing of different experiments.

The computing time was approximately ~1.25min for each data set with a Pentium IV 3GHz and 1GB of RAM using Excel 2002 (Office XP) with WinXP.

The final results of the calculations are shown in comparison to the results published by Helmann et al., 2003, in Table 2. The average accuracy was 93.3%  $\pm$ 4.2%. Three genes are missing in the comparison: the gene *zosA* could not be found in the raw data files and for the genes *yybl* and *ytpQ* no differential expression according to the 3-fold cutoff could be detected.



## **2.2 BRB-ArrayTools**

The BRB-ArrayTools 3.3 is an Excel add-in that provides a number of tools for the computation and visualization of, e.g., SAM or Cluster analysis. Normalized expression ratios are calculated from the microarray raw data in a similar way to MADA. The program including a manual can be downloaded at the web page of the National Cancer Institute, Division of Cancer Treatment and Diagnosis, Biometric Research Branch at <http://linus.nci.nih.gov/BRB-ArrayTools.html>.

### **2.2.1 Based on the data set from Peplies et al., 2004**

We have tried to import the raw data using the 'collate data' import wizard in the single channel mode. The original QuantArray file could not be imported since the header contained more than 50 rows. Therefore, the header was removed and gene names and signal intensities of the channel 2 data were imported (the background correction function is not accessible in this mode). However, since we could not provide reasonable data for channel 1 (it contained data for an additionally hybridized control target whose corresponding capture probe represents only a very small part of the probe set), no calculation results were obtained. In principle, the tool is able to import raw data coming from a single channel but since it is exclusively made for gene expression analysis, reasonable data coming from two channels are required to calculate the expression ratios.

### **2.2.2 Based on the data set from Helmann et al., 2003**

The six raw data files provided by Helmann et al., 2003, were initially placed in a separate folder. An experiment description file had then to be created. It is required by the BRB-ArrayTools to identify data sets in which the fluorescent dyes have been swapped. An Excel file was set up with the six raw data file names in the first column and an identifier in the second column. Experiments "without" the dye swap were identified by '0' and files with the dye swap by '1'.

The 'collate data', 'data import wizard' was then used with channel 1 set to 'green' and channel 2 set to 'red'. The background adjustment was activated. Calculation was done without spot or gene filtering and the 'Median of entire array' normalization.

The computing time was approximately ~0.1min with a Pentium IV 3GHz and 1GB of RAM using Excel 2002 (Office XP) with WinXP.

The final results of the calculations are shown in comparison to the results published by Helmann et al., 2003, in Table 3. The average accuracy was 94.4%  $\pm$ 4.4%. The log expression values  $\geq 5$  and  $\leq -1$

which can occur due to, e.g., signals not significantly above the background in one of the two channels where manually removed.

Three genes are missing in the comparison: the gene *zosA* could not be found in the raw data files and for the genes *yybI* and *ytpQ* no differential expression according to the 3-fold cutoff could be detected.

Helmann et al. 2003, Tab.1		BRB-ArrayTools						Transformed to foldinduction						Average	Accuracy [%]	
Gene(s)	Foldinduction	Normalized log-expression						Dye swap						Average	Accuracy [%]	
		Experiment 25663	Experiment 25664	Experiment 25665	Experiment 25666	Experiment 25667	Experiment 25668	Experiment 25663	Experiment 25664	Experiment 25665	Experiment 25666	Experiment 25667	Experiment 25668			
comER	9.9	3.81	3.36	4.04	3.47	2.92	3.01	10.25	16.40	11.09	7.55	8.05	13.99	11.22	88.2	
yybF	9.81	3.45	3.35	3.41	3.29	3.13	3.27	10.94	10.18	10.66	9.75	8.75	9.66	9.99	98.2	
mrgA	9.64	3.33	3.36	3.34	3.16	3.26	3.31	10.06	10.29	10.11	8.95	9.55	9.89	9.81	98.3	
ywfM	8.94	3.63	3.51	3.48	3.82	2.96	2.84	12.36	11.40	11.19	14.17	7.80	7.15	10.68	83.7	
comEA	8.27	3.01	3.04	2.96	3.18	3.35	3.24	8.05	8.21	7.79	9.05	10.17	9.45	8.79	94.1	
comGA		3.04	3.25	3.16	3.64	3.42	3.38	8.22	9.52	8.91	12.45	10.73	10.43	10.04		
comGB		3.22	2.97	3.00	3.11	3.11	3.83	9.31	7.85	8.03	8.66	8.63	14.18	9.44		
comGC		2.90	2.85	2.81	3.34	3.06	3.18	7.46	7.19	7.03	10.13	8.36	9.04	8.20		
comGD		3.22	3.28	3.14	3.05	3.06	3.90	9.32	9.69	8.81	8.29	8.35	14.92	9.90		
comGE		2.93	2.96	2.90	3.37	3.29	3.61	7.63	7.76	7.44	10.32	9.77	12.21	9.19		
comGF		2.89	3.02	2.85	2.85	2.80	3.50	7.41	8.13	7.22	7.22	6.97	11.33	8.05		
comGG		2.02	2.03	2.01	2.32	2.18	2.26	4.05	4.08	4.03	4.99	4.54	4.77	4.41		
comG(ABCDEFG)	7.93 ± 1.74													8.46 ± 1.95	93.7	
meiA	7.31	2.94	2.91	2.98	3.16	2.94	3.06	7.69	7.53	7.91	8.95	7.67	8.32	8.01	91.2	
cwiJ	7.64	3.08	3.15	2.94	2.81	2.77		8.44	8.88	7.68	7.01	6.83		7.77	98.3	
gbsA		3.11	4.64	3.07	2.89	2.82	3.32	8.62	24.99	8.40	7.41	7.06	10.01	11.08		
gbsB		2.66	2.64	2.61	2.49	2.71	5.00	6.30	6.23	6.12	5.62	6.55		6.16		
gbsAB	7.62 ± 0.52													8.62 ± 3.48	88.4	
msmR		2.30	2.31	2.36	2.42	2.29	2.29	4.93	4.97	5.14	5.34	4.90	4.89	5.03		
msmE		3.31	3.32	3.40	2.81	2.81	2.85	9.94	9.97	10.57	7.01	7.02	7.19	8.62		
amyD		3.04	2.95	3.07	3.11	2.98	2.94	8.24	7.72	8.37	8.61	7.91	7.65	8.08		
amyC		3.33	3.26	3.37	2.83	2.91	2.88	10.05	9.58	10.34	7.13	7.49	7.38	8.66		
msmRE amyDC	7.44 ± 1.67													7.60 ± 1.73	97.9	
dppA		2.55	2.63	2.59	2.69	2.65	3.13	5.86	6.18	6.00	6.45	6.26	8.75	6.58		
dppB		2.91	4.53	2.81			3.16	4.05	7.50	23.16	7.03	1.00	8.96	16.55	10.70	
dppC		2.75	2.60	2.76	3.05	2.91	3.71	6.74	6.04	6.76	8.30	7.49	13.08	8.07		
dppD		2.61	2.58	2.46	3.06	3.04	3.11	6.10	5.97	5.49	8.34	8.20	8.66	7.13		
dppE		2.79	2.73	2.86	2.63	2.54	3.07	6.93	6.63	7.27	6.20	5.83	8.40	6.88		
dppABCDE	7.04 ± 0.81													7.87 ± 1.68	89.4	
zosA	5.97	not available in the raw data files														
appD		2.63	2.59	2.70	2.45	2.38	2.69	6.21	6.02	6.52	5.48	5.21	6.45	5.98		
appF		2.46	2.32	2.44	2.51	2.45	2.45	5.51	4.99	5.41	5.70	5.48	5.45	5.42		
appDF	5.53 ± 0.34													5.70 ± 0.40	97.0	
kata	5.34	2.52	2.51	2.54	2.27	2.29	2.47	5.72	5.70	5.81	4.81	4.88	5.55	5.41	98.7	
livB		3.03	3.04	2.98	2.63	2.68	2.99	8.18	8.25	7.91	6.17	6.42	7.96	7.48		
livC		2.80	2.27	2.70	1.88	1.86	2.12	6.95	4.81	6.48	3.68	3.62	4.35	4.98		
leuB		10.56	2.14	2.11	1.89	2.02	2.46		4.42	4.32	3.71	4.04	5.52	4.40		
leuC		2.06	2.13	2.22	2.32	2.24	2.29	4.16	4.37	4.65	4.99	4.73	4.88	4.63		
livBC leuBC	5.01 ± 0.71													5.37 ± 1.42	93.2	
ureA		2.75	2.55	2.40	2.40	2.46		6.71	5.86	5.26	5.27	5.51		5.72		
ureB		2.43	2.30	2.29	2.89	2.15	2.84	5.38	4.92	4.90	7.42	4.45	7.17	5.71		
ureC		1.88	2.14	1.81	1.89	1.84	1.76	3.68	4.41	3.51	3.70	3.57	3.38	3.71		
ureABC	4.96 ± 1.4													5.05 ± 1.16	98.3	
yckD		2.49	2.09	2.29	2.02	1.87	3.85	5.60	4.26	4.89	4.06	3.66	14.37	6.14		
yckE		2.40	2.24	2.34	2.49	2.37	2.36	5.28	4.73	5.07	5.63	5.16	5.14	5.17		
yckDE	4.96 ± 0.1													5.65 ± 0.69	87.7	
yybI	4.92	0.59	0.52	0.50	-0.20	0.21	0.13	1.50	1.44	1.41	0.87	1.16	1.09	not expressed		
appB		2.08	1.92	2.02	2.26	2.19	2.15	4.24	3.79	4.04	4.79	4.56	4.45	4.31		
appC		2.45	2.53	2.40	2.24	2.37	2.89	5.46	5.76	5.29	4.73	5.15	7.43	5.64		
appBC	4.89 ± 0.9													4.98 ± 0.94	98.3	
yxbC	4.6	2.45	2.28	2.35	2.03	1.96	2.47	5.48	4.87	5.08	4.07	3.89	5.55	4.82	95.4	
yxbB		2.28	2.16	2.14	2.30	2.07	2.25	4.84	4.47	4.42	4.92	4.19	4.74	4.60		
yxbA		2.22	2.14	2.29	1.85	1.97	3.24	4.65	4.42	4.89	3.60	3.93	9.42	5.15		
yxbBA	4.5 ± 0.1													4.87 ± 0.39	92.3	
yufN	4.44	2.26	2.40	2.34	2.12	1.92	2.95	4.79	5.27	5.07	4.34	3.78	7.71	5.16	86.0	
yvaW		1.69	1.64	1.71	1.71	1.64	1.01	3.23	3.12	3.27	3.28	3.11	2.01	3.00		
yvaY		2.57	2.47	2.60	2.62	2.51	2.54	5.95	5.55	6.08	6.16	5.71	5.82	5.88		
yvaWY	4.36 ± 0.2													4.44 ± 2.03	98.2	
yhdG	4.27	2.52	1.74	1.99	2.53	2.14		5.72	3.35	3.98	5.80	4.41		4.65	91.8	
slpT	4.25	2.21	1.90	2.16	2.31	2.08	2.65	4.63	3.74	4.47	4.96	4.23	6.30	4.72	90.0	
ygbA	4.18	2.07	2.17	2.08	1.98	1.95	2.27	4.20	4.49	4.24	3.94	3.86	4.83	4.26	98.1	
ppsA		1.67	2.44	1.48	1.71	1.89	1.91	3.18	5.41	2.80	3.28	3.72	3.76	3.69		
ppsB		-1.10	0.67	0.75	1.06	1.01	1.95		1.59	1.68	2.09	2.02	3.87	2.25		
ppsC		2.03	2.07	2.12	1.93	2.12	3.23	4.09	4.19	4.36	3.81	4.35	9.36	5.03		
ppsD		2.43	1.10	1.03	-0.51	0.75	2.25	5.39	2.15	2.04	0.70	1.68	4.75	2.78		
ppsE		2.09	2.18	2.13	1.87	1.82	1.99	4.27	4.53	4.37	3.64	3.52	3.98	4.05		
ppsABCDE	3.93 ± 0.52													3.66 ± 1.09	90.6	
yuiA	3.9	2.02	2.47	1.97	2.33	1.97	2.36	4.04	5.55	3.93	5.03	3.93	5.13	4.60	84.7	
ykvK		1.70	1.85	1.65	1.74	1.61	2.17	3.25	3.60	3.13	3.35	3.04	4.51	3.48		
ykvL		2.31	2.32	2.33	1.99	2.02	2.17	4.98	5.00	5.01	3.96	4.05	4.50	4.58		
ykvM		1.89	1.94	1.89	1.93	1.80	1.99	3.71	3.85	3.70	3.81	3.49	3.97	3.75		
ykvKLM	3.83 ± 0.6													3.94 ± 0.57	97.2	
yheK	3.76	2.03	1.99	2.03	1.90	1.78	1.95	4.09	3.99	4.10	3.74	3.43	3.87	3.87	97.2	
yisS	3.67	1.80	1.71	1.82	2.05	1.96	2.07	3.48	3.27	3.54	4.14	3.90	4.19	3.75	97.7	
ykfB		2.00	1.92	1.87	2.10	1.84	2.74	4.01	3.79	3.66	4.28	3.57	6.66	4.33		
ykfC		1.69	1.58	1.63	2.02	1.91	2.06	3.22	2.99	3.10	4.04	3.76	4.16	3.55		
ykfD		1.83	1.76	1.75	1.80	1.72	2.23	3.55	3.40	3.36	3.47	3.30	4.70	3.63		
ykfBCD	3.67 ± 0.34													3.84 ± 0.43	95.7	
ywcE	3.5	1.77	1.63	1.82	2.06	1.82	1.90	3.41	3.10	3.54	4.17	3.54	3.73	3.58	97.7	
yocB	3.29	1.76	1.81	1.83	1.56	1.63	1.81	3.38	3.50	3.56	2.95	3.10	3.51	3.33	98.7	
yhdC	3.27	1.83	1.80	1.65	1.74	1.73	1.71	3.54	3.48	3.15	3.34	3.31	3.26	3.35	97.7	
bioA		1.78	1.80	1.78	1.72	1.73	1.74	3.44	3.47	3.44	3.29	3.32	3.35	3.39		
bioF		1.51	1.45	1.50	2.01	1.91	1.98	2.85	2.73	2.83	4.01	3.76	3.96	3.36		
bioD		1.81	1.75	1.83	1.67	1.70	1.76	3.51	3.36	3.55	3.19	3.24	3.39	3.37		
bioB		1.52	1.56	1.48	1.94	1.72	1.79	2.87	2.94	2.79	3.84	3.30	3.46	3.20		
bioAFDB	3.26 ± 0.11													3.33 ± 0.09	97.9	
rapH	3.21	1.51	1.49	1.58	1.93	1.79	1.89	2.85	2.82	2.99	3.81	3.47	3.70	3.27	98.1	
yybG	3.14	1.83	1.86	1.73	1.41	1.56	1.63	3.55	3.63	3.31	2.66	2.96	3.10	3.20	98.1	
opuBB	3.14	1.76	2.06	1.85	1.45	1.53		3.39	4.18	3.60	2.74	2.90		3.36	93.5	
ytpQ	3.1	0.02	0.04	0.08	-0.09	-0.10	-0.03	1.01	1.03	1.06	0.94	0.93	0.98	not expressed		
gltB	3.07	1.80	1.84	1.78	1.35	1.43	1.63	3.47	3.59	3.43	2.54	2.6				

### 3 Conclusions

The functionality of MADA could be demonstrated by processing a raw data set published by Peplies et al. in 2004 and representing a completely defined experiment in the field of organism identification. The corresponding results, originally obtained by manual data processing, could be reproduced with an accuracy of 100%.

The functionality and effectiveness of the MADA outlier was demonstrated with a manually modified raw data set taken from the study of Peplies et al., 2004.

MADA was able to reproduce results from Helmann et al., 2003, with an accuracy of 97.4%  $\pm$ 2.4% which is in the range or even better compared to similar software tools such as MarC-V (93,3%  $\pm$ 4.2%) or BRB-ArrayTools (94.4%  $\pm$ 4.4%). The observed variation is presumably based on differences in the underlying calculation and normalization procedures.

Compared to other Microsoft Excel-based microarray data analysis tools, MADA is able to process raw data from experiments in both fields of application, identification and gene expression profiling.

The speed of data processing for MADA is in the range of the other tools tested. However, it is planned to revise the algorithms of the next MADA version to significantly optimize the calculation speed.

**Update 02.2007: The performance of MADA 2.0 was optimized. Compared to previous versions of MADA an up to 8 times faster calculation was achieved on average.**

Compared to the other tools tested, MADA allows the import of various data formats using a number of predefined import filters. Additional formats can easily be imported by setting up specific import filters using MADA's 'Edit filter' tool.

Compare to the other tools tested, MADA provides a batch modus for the simple parallel processing of various independent data sets.

Additional general outstanding features of MADA are an intuitive user guidance, high transparency during data processing, the possibility to recalculate/reprocess data on each level of analysis without restarting the complete procedure, and the availability of a comprehensive manual.

#### **4 References**

**J. Peplies, S. C. K. Lau, J. Pernthaler, R. Amann, and F. O. Glöckner.** 2004. Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. *Environmental Microbiology* 6:638-645.

**J.D. Helmann, M.F. Wu, A. Gaballa, P.A. Kobel, M.M. Morshedi, P. Fawcett and C. Paddon.** 2003. The global transcriptional response of *Bacillus subtilis* to peroxide stress is coordinated by three transcription factors. *J. Bacteriol.* 185(1):243-53.

**J.J. Schageman, M. Basit, T.D. Gallardo, H.R. Garner and R.V. Shohet.** 2002. MarC-V: A spreadsheet-based tool for analysis, normalization, and visualization of single cDNA microarray experiments. *BioTechniques* 32:338-344.

**BRB-Array Tools.** National Cancer Institute, Division of Cancer Treatment and Diagnosis, Biometric Research Branch. <http://linus.nci.nih.gov/BRB-ArrayTools.html>.