

Micro Array Data Analyser

>>> Manual <<<

The screenshot displays the MicroArrayDataAnalyser 2.0 interface. On the left, a Notepad window shows raw data file information and a table of gene statistics. The main window features a central control panel with buttons for 'Import Raw Data', 'Extract Data', 'Calculate', 'Plot Chart', and 'Save Data'. A flowchart on the right details the calculation steps: 'Local background correction' (Background Corrected Spot Intensity = Spot Intensity - Spot Background Intensity), 'Is Signal?' - Signal significance test (Spot Intensity > Background Intensity + 2 * Background StDev), 'Outlier' test, and 'Mean of background corrected replicates (where signal is significant and no outlier)'. Below the flowchart, a 'START' button is visible. At the bottom, a table lists gene names and their corresponding data values, and a bar chart shows the normalized mean of replicates for each gene.

ID	Name	Ch1 Median	Ch1 Mean	Ch1 SD	Ch1 B Median	Ch1 B Mean
1	gen2627/172	6346	6693	756.96	3867	4162
2	gen2627/172	6425	6694	645.5	4006	4287
3	gen2627/172	6182	6457	536.11	3964	4177
4	gen2627/172	6283	6513	768.73	4816	4327
5	gen2627/172	6450	6729	782.5	4834	4847
6	gen2627/688	6760	7824	759.87	3769	4837
7	gen2627/688	3914	4156	352.2	3682	4878
8	gen2627/688	3944	4171	998.75	3559	3922
9	gen2627/688	3658	3940	156.85	3648	3844
10	gen2627/688	3616	392			

Gene	115.77	11.81	53.82	103.98
GeneA	122.74	26.48	79.13	96.26
GeneA	336.96	4.11	32.26	331.24
GeneB	1611.19	33.45	119.46	1577.74
GeneB	1670.26	62.63	210.84	1617.63
GeneB	1739.20	2.36	24.47	1736.86
GeneC	1406.84	26.76	87.27	1380.07
GeneC	1369.66	6.97	41.75	1362.69
GeneC	1376.65	14.08	66.76	1362.57

Gene	115.77	11.81	53.82	103.98	122.74	26.48	79.13	96.26	336.96	4.11	32.26	331.24	1611.19	33.45	119.46	1577.74	1670.26	62.63	210.84	1617.63	1739.20	2.36	24.47	1736.86	1597.69	28.20815433	1406.84	26.76	87.27	1380.07	1369.66	6.97	41.75	1362.69	1376.65	14.08	66.76	1362.57	1365.11	138821176
------	--------	-------	-------	--------	--------	-------	-------	-------	--------	------	-------	--------	---------	-------	--------	---------	---------	-------	--------	---------	---------	------	-------	---------	---------	-------------	---------	-------	-------	---------	---------	------	-------	---------	---------	-------	-------	---------	---------	-----------

replicates 1
Normalized(% of mean of #genes)[Ch2 Mean of Replicates of BC Mean]

Andreas Ellrott, Jörg Peplies

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

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

Reference:

Ellrott A, Würdemann C, Peplies J, Glöckner FO:
MADA - A software tool for the highly transparent
initial processing of microarray raw data based on
Microsoft Excel.
submitted.

(c) 05.2007

OVERVIEW	7
1 IMPORT RAW DATA.....	9
2 EXTRACT DATA	13
2.1 Extraction filter.....	15
2.1.1 Commands available for setting up an extraction filter	16
2.1.2 Example 1	19
2.1.3 Example 2, Filter for ScanArray file.....	20
2.1.4 Example 3, Filter for GenePix file.....	21
3 CALCULATE	23
3.1 Local background correction.....	26
3.2 Signal significance test.....	26
3.3 Outlier test.....	28
3.4 Mean of Replicates.....	29
3.4.1 Normalisation	30
3.4.1.1 Mean / Median of overall signal.....	30
3.4.1.2 Mean of corresponding replicates.....	31
3.4.1.3 Mean of selected genes.....	31
3.4.1.4 Mean / Median of overall background signal	32
3.4.1.5 Lowess normalisation	32
3.5 Ratio & intensity	34
4 PLOT CHART	35
4.1 Mean of replicates.....	36
4.2 Mean of replicates Chy over Mean of replicates Chx.....	38
4.3 Ratio over Intensity	40
4.4 Ratio over Ratio.....	42
5 SAVE DATA	45
6 REMOVE DATA SHEETS	47
7 SYSTEM LOG	49
8 INSTALLATION AND SETUP	51
8.1 Installing MADA on your PC	51
8.2 Setup Excel to run MADA	52
9 SYSTEM REQUIREMENTS AND PERFORMANCE.....	55
10 TROUBLESHOOTING.....	59
APPENDIX.....	67

This page is intentionally left blank to support double-sided printouts.

Due to its high parallelism, microarray analysis produces comprehensive sets of raw data encompassing various parameters describing the hybridisation results for a large number of spotted capture probes. Sophisticated data analysis including various calculation steps is necessary in order to generate reliable results. Only computer programs are able to handle these processes within a reasonable time period.

MADA, just another Micro-Array Data Analyser?

Various theoretical approaches as well as specific software tools are available for microarray data analysis. While some of these tools apply specific mathematic approaches to solve certain problems, others attempt to cover the full range of known methods.

An integral aspect of scientific work is rendering processes transparent, so as to unveil their complexity and to allow for a thorough understanding of how things operate. Unfortunately, many tools tend to be as complex as the problems they address, and these tools often adopt a black box approach, in which only results are shown but the calculation steps are hidden.

MADA does not encompass new approaches for data calculation, and also does not apply the full range of methods currently available. Instead, its main objective is to make data analysis more transparent and to provide complete control over each and every calculation step.

MADA's intuitive user interface allows common methods of microarray data analysis to be performed in a transparent, flexible, and natural manner. Inexperienced users will find easy access to important microarray data analysis commands, while advanced users are provided with the flexibility and full-control they sometimes require. Implemented in Microsoft Excel, MADA is part of a powerful and well-known software environment.

Microsoft Excel, a good choice for programming?

Many computer scientists would never consider writing a program with Microsoft Excel, so why have we? Most researchers are so familiar with the software that if ever you ask them how they did their calculations, charts, and graphs, an exceedingly common answer is Excel. It is therefore also practical to use Excel's software environment for automatic microarray data analysis. As a positive side effect, users have direct access to the numerous Excel tools. Another clear advantage is the transparency of calculation provided by the program. Whenever possible, MADA will not only calculate a value, but will provide its underlying formula, so that the user can trace back the calculation process. Finally, Excel offers the ability to present data in concise user-defined charts and graphs, some of which are directly implemented in MADA - and with Excel just a mouse-click away, the user is completely free to adjust them.

Although concerns cited against Excel are often of a philosophical nature, it does have some disadvantages as a software programming environment. However, our focus was clearly on maximising usability. We have tried our very best to make MADA as stable as possible, but due to the large variety of different Excel versions and operating systems available, we can not guarantee the complete absence of bugs. If you do encounter any problems, please report them to aellrott@mpi-bremen.de, and we will attempt to correct them.

Software Validation

The functionality and accuracy of MADA was evaluated against published reference data sets. Detailed information about the used raw data sets, the performed calculation procedures and the obtained results can be found at <http://www.megx.net/mada>.

Additionally, the validation and the used raw data sets could serve as guidance and reference for an example Project.

Disclaimer

We have to point out that the use of MADA and all its components, including this manual, is on your own risk. The authors will not guarantee for the correctness of any calculations and are not liable for any damages arising from the use of the software. You have to accept the 'freeware licences agreement' if you use MADA. A copy of this agreement can be found at the end of this manual. The current version is always supplied together with the corresponding program.

Overview

MADA's highly structured design directly reflects its approach to data analysis. Such an analysis can easily be realised by simply following the corresponding buttons in the main window in a step-by-step procedure (Fig. 1). This design not only provides unencumbered access, it also allows one to jump back and restart recalculations at any previous step. Therefore, different calculation parameters and/or methods can be tested in a fast and efficient way. For a comprehensive overview compare Fig. 2.

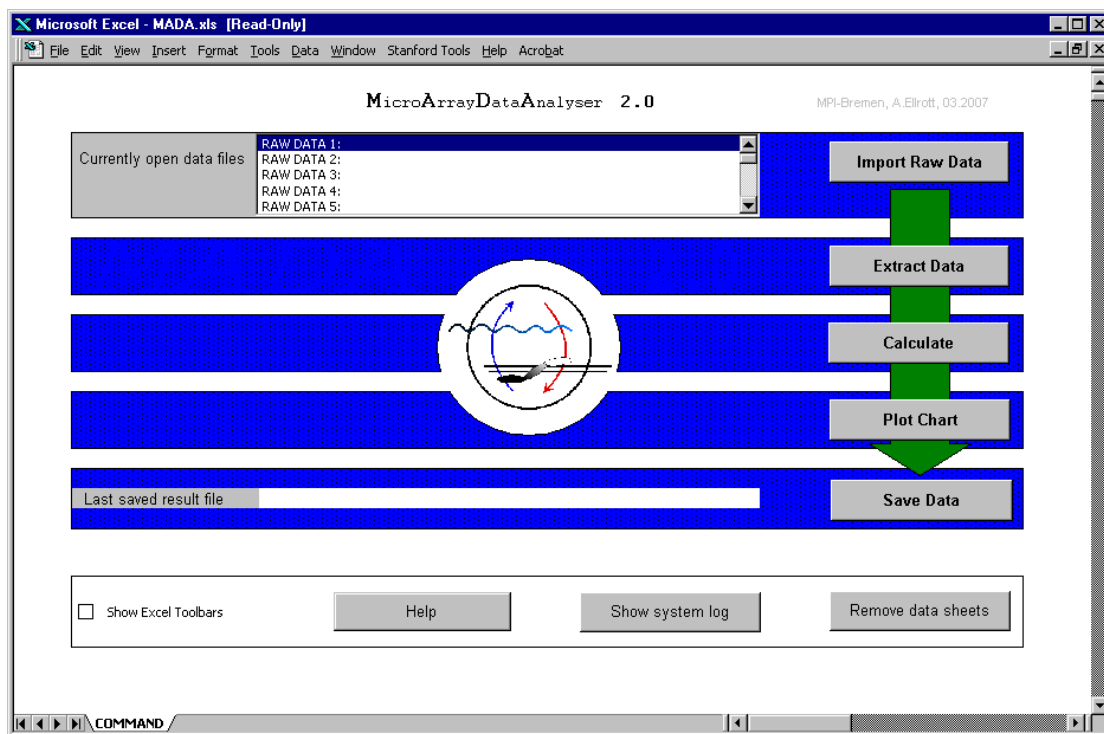


Figure 1: MADA's main window

Data Import

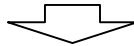
Import of up to **24** data files (excel or delimited text files).

Each file can hold the information of max. 60,000 spots and 3 different channels.

**Data Extraction**

Extraction of the data of interest from the raw data sets:

- Gene name / ID
- Chx spot intensity
- Chx spot background intensity
- Chx spot background standard deviation
- Chx wavelength or fluorophore name (optional)

**Calculation****Local background correction**

Chx background corrected spot intensity =
Chx spot intensity – Chx spot background intensity

Signal significance test

Chx spot intensity >
Chx background intensity + t * Chx background StDev
t (default) = 2

**Outlier test**

Tests the remaining replicates for the presence of outliers.

**Mean of replicates**

Probe signal intensities of the filtered data set are calculated by combination of probe replicates (arithmetic mean).

Normalization

A normalization factor can be calculated based on:

- Mean or median of overall spot signal intensities
- Mean of corresponding replicates
- Mean of selected genes
- Mean or median of overall background signal intensities

**Ratio and intensity**

ratio = $\log_2 (A / B)$

intensity = $\log_{10} (A * B)$

A, B = (normalized) mean of replicates of channel 1, 2 or 3

Normalization

Lowess regression on ratio values with selectable smooth factor.

**Plot Chart**

Visualization of results via predefined charts, such as R-I-plots.

**Save Data**

Raw data, result data and charts can be saved in a new excel workbook.

A export file can be created combining selected data from different data sets.

Chx: Channel 1 to 3 [red]: Changes in MADA 2.0

Figure 2: General scheme of MADA

1 Import Raw Data

This chapter explains how to import microarray raw data into MADA.

Since MADA is not an image analysis software you can NOT import images of scanned microarrays. Initial image analysis for quantification of spot and background intensities has to be done by external programs. Mostly, they are included in the microarray scanner software.

Starting point of data analysis using MADA is a raw data file in which each spot is described by a set of parameters. This can be an Excel worksheet or a delimited text file as long as the data are arranged in a columnar list in which each line represents a single spot of the array.

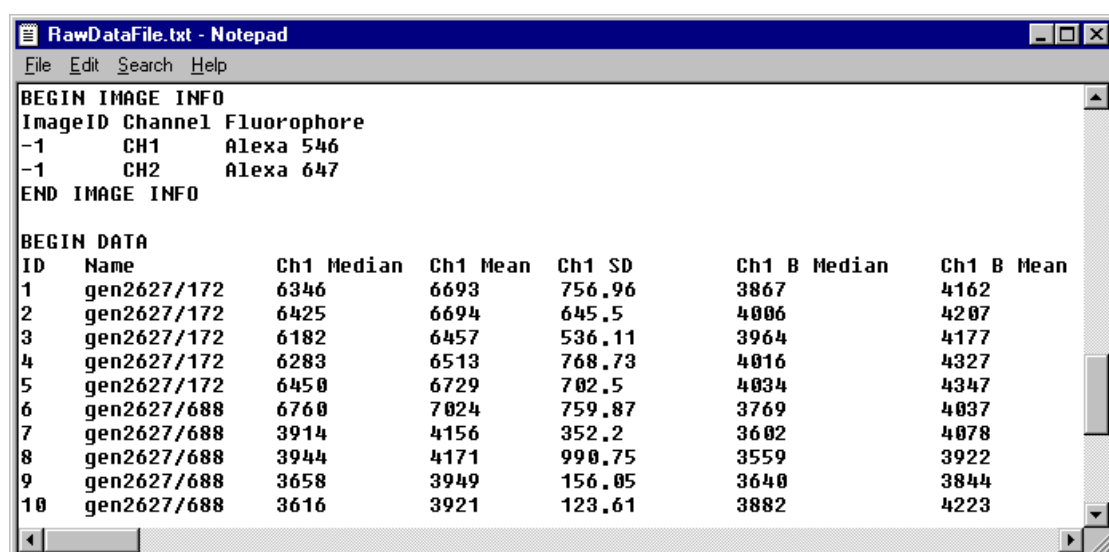
Most microarray image analysis software tools provide a large variety of spot-describing parameters but for common data analysis as implemented in MADA, basically only 4 different parameters are of interest.

Make sure that the raw data file contains the following information:

- Gene name
- Chx spot intensity
- Chx spot background intensity
- Chx spot background standard deviation

Chx: Channel 1 to 3

For better overview, a fluorophore name or wavelength should be assigned to each of the scanner channels represented in the data file.



```
RawDataFile.txt - Notepad
File Edit Search Help
BEGIN IMAGE INFO
ImageID Channel Fluorophore
-1 CH1 Alexa 546
-1 CH2 Alexa 647
END IMAGE INFO

BEGIN DATA
ID Name Ch1 Median Ch1 Mean Ch1 SD Ch1 B Median Ch1 B Mean
1 gen2627/172 6346 6693 756.96 3867 4162
2 gen2627/172 6425 6694 645.5 4006 4207
3 gen2627/172 6182 6457 536.11 3964 4177
4 gen2627/172 6283 6513 768.73 4016 4327
5 gen2627/172 6450 6729 702.5 4034 4347
6 gen2627/688 6760 7024 759.87 3769 4037
7 gen2627/688 3914 4156 352.2 3602 4078
8 gen2627/688 3944 4171 990.75 3559 3922
9 gen2627/688 3658 3949 156.05 3640 3844
10 gen2627/688 3616 3921 123.61 3882 4223
```

Figure 3: Organization of a typical raw data file

Note: Excel can misinterpret identifiers, e.g., the gene name DEC-1 is interpreted as date and automatically changed to 1-Dec during import [Zeeberg et al. 2004].

Procedure of data import:

a) Press the 'Import Raw Data' button in the main window and a window called 'Select files for import' will open.

b) Define your working directory by typing in the path (this can also be done in the MADA main window) or select it via the 'BROWSE' button. The working directory is the standard folder from which raw data files are imported and results are stored after data processing.

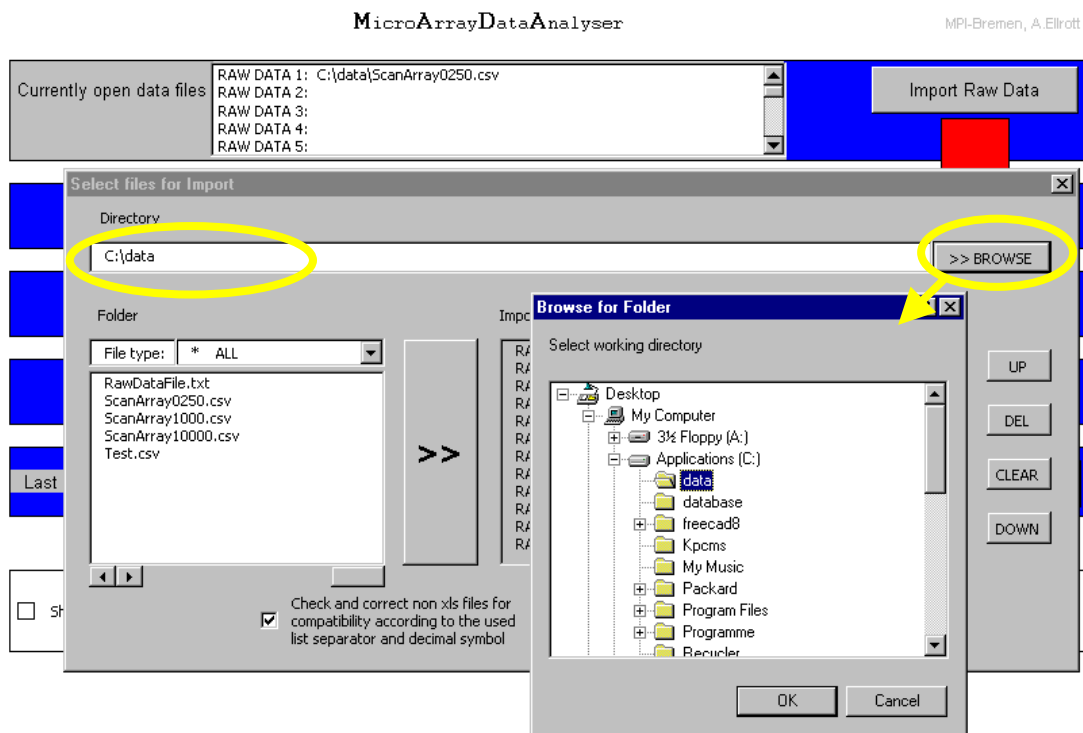


Figure 4: Definition/selection of the working directory

c) Files within the working directory are listed on the left according to the file type chosen. Select the raw data files of interest and transfer them to the right box via the transfer button. By using the 'UP', 'DOWN', 'DEL' or 'CLEAR' functions the files can assigned to one of MADA's twenty-four 'RAW DATA' sheets. 'START' the data import.

Note: 'RAW DATA' sheets already in use are shown in the 'Currently open data files' list of the MADA main window. If a new file is assigned to a 'RAW DATA' worksheet which is already in use, it will be overwritten.

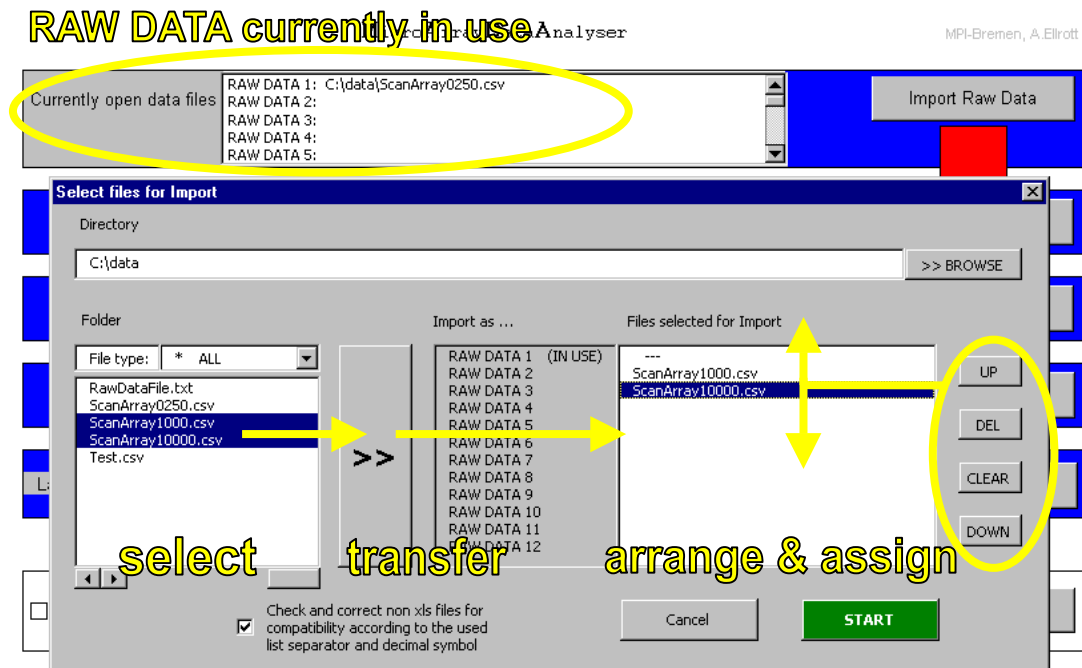


Figure 5: Selection, transfer and assignment of raw data files

Note: Sometimes, Excel can not correctly interpret data within delimited files because the list delimiter and the decimal separator do not correspond to the presets of the operating system. In this case import or calculation can fail since, e.g., the German number 1,5 is written 1.5 on a US system.

MADA has a build in option to check the compatibility of non xls files during the 'transfer' step, if the option 'Check and correct non xls files ...' is selected. If any incompatibility is found MADA will give a message and asks the user which action to take. We recommend the '>> Convert file before import' –option, which will automatically create a new compatible file copy indicated by the word CONVERTED as prefix of the filename. This file will be used for import afterwards and is stored at the same location like the original file. The original file will stay untouched.

For further information please refer to the *Troubleshooting* section.

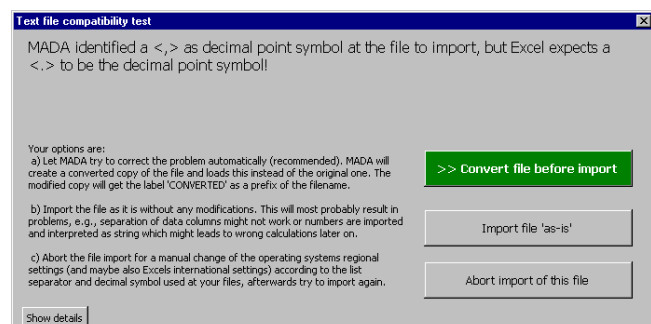


Figure 6: File compatibility check

d) In case of importing data from Excel files encompassing more than one worksheet, you will be asked to select the appropriate raw data sheet.

Make your selection and press 'OK' to continue.

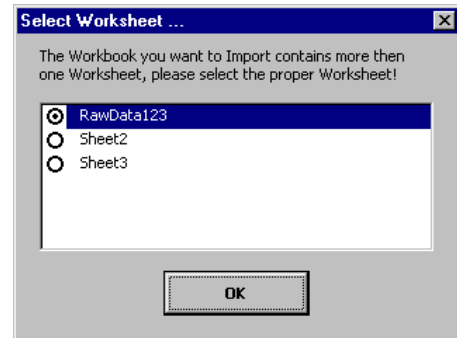


Figure 7: Selection of working sheet

e) The progress bar will display the current status of the data import.

A message is shown when the import is done.

Afterwards, the program will automatically switch back to the main window of MADA.

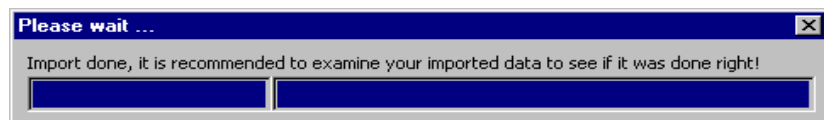


Figure 8: Progress of data import

Note: Always check the 'RAW DATA' worksheets for correctness. If the data does not look like as you have expected, refer to the *Troubleshooting* section.

2 Extract Data

This chapter explains how to use the 'Extract Data' function. It is extracting the parameters describing a spot (listed in the 'Import Raw Data' section Gene name, Chx spot intensity, etc.) from the 'RAW DATA' worksheets and writes them to the corresponding 'RESULT DATA' worksheets for further calculation and analysis.

The 'Extract Data' button within the main window of MADA will open the corresponding dialog window.

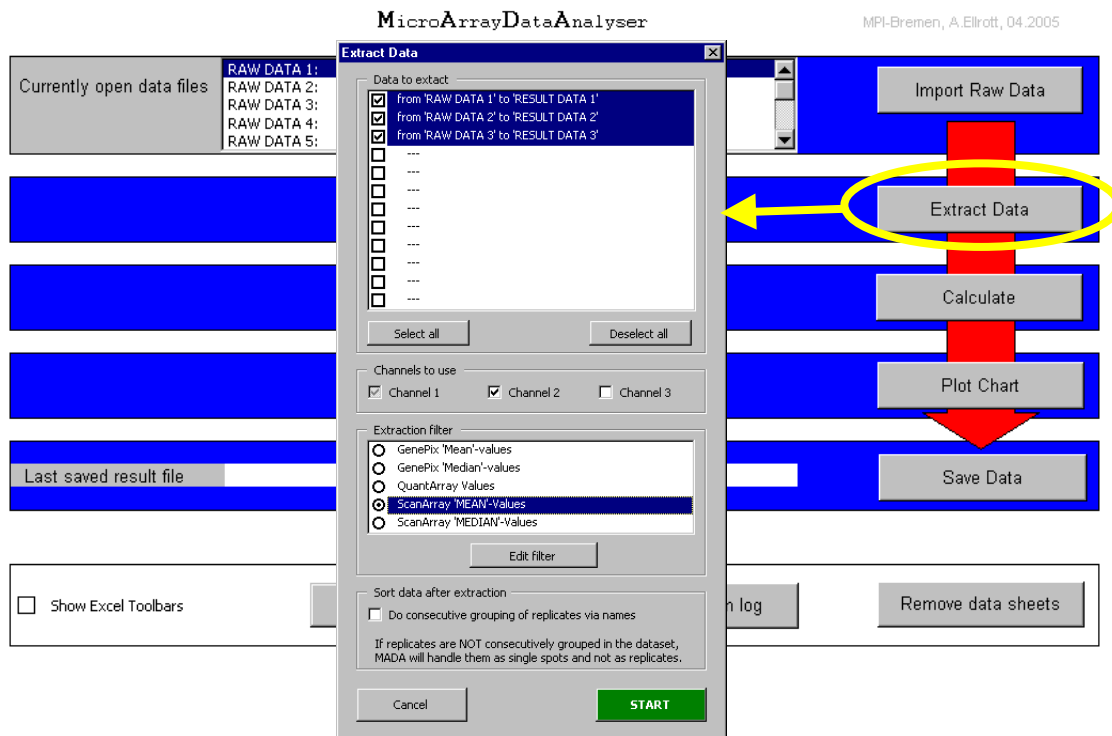


Figure 9: Main dialog window of the 'Extract Data' function

- a) Select the 'RAW DATA' worksheets from which the data should be extracted.
- b) Select the channels (fluorophores) of interest.
Note: Of course, at least one channel is required for data analysis. Therefore, you can not deselect the first one. If you select multiple channels, make sure that they are present in the 'RAW DATA' worksheets.
- c) Chose an extraction filter that corresponds to the structure of your raw data sets. The extraction filter is like a map that guides MADA to the locations in the data sets where the information of interest are located. If there is no extraction filter fitting to your data, you must create an own one as described in the section 'Extraction filter'.

- d) If probes are spotted in replicates on your microarray, all of these replicates must be named identically and consecutively grouped in the data file. If the replicates are spreading within the data set, mark the box 'Do consecutive grouping of replicates via names' to avoid replicates of being handled as single spots during later calculations.
- Note:** Because the 'fast' sort method of Excel is used, consecutive grouping will always lead to an alphabetical order of the probe names in the data set.

Press 'START' to begin the extraction procedure. A progress bar will inform you about the current status.

In case of errors, first check if the correct extraction filter was selected and make sure that the selected channels exist. Further information can be found in the 'troubleshooting' section of this manual.

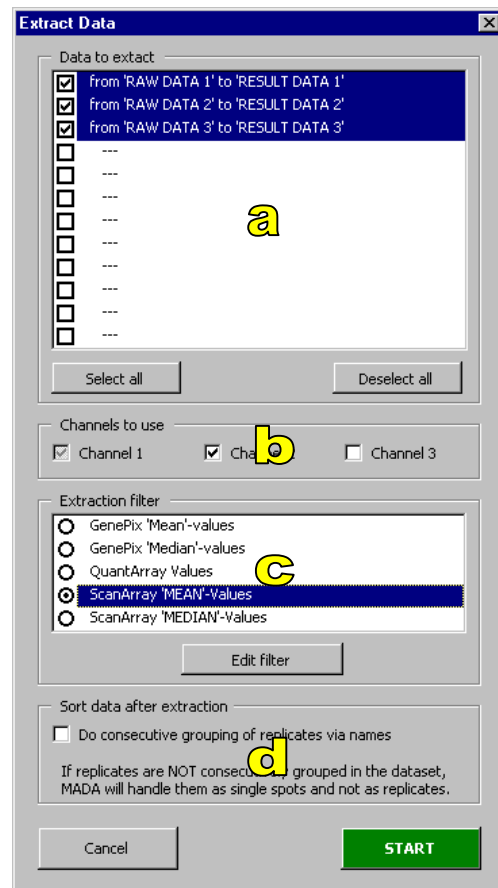


Figure 10: Example for settings of the 'Extract Data' function

Begin Data						
Number	Row	Column	Name	ch1 Intensity	ch1 Background	ch1
1	1	1	A	19723.32617		0
2	1	2	B	30836.73438		0
3	1	3	C	28522.06055		0
4	1	4	A	18412.08203		0
5	1	5	B	31124.67383		0
6	1	6	C	26109.67383		0
7	1	7	A	16774.34766		0
8	1	8	B	32776.30664		0
9	1	9	C	27586.77539		0

No consecutive distribution of probe replicates on the microarray and within the data file.

Begin Data						
Number	Row	Column	Name	ch1 Intensity	ch1 Background	ch1
1	1	1	A	19723.32617		0
4	1	4	A	18412.08203		0
7	1	7	A	16774.34766		0
2	1	2	B	30836.73438		0
5	1	5	B	31124.67383		0
8	1	8	B	32776.30664		0
3	1	3	C	28522.06055		0
6	1	6	C	26109.67383		0
9	1	9	C	27586.77539		0

Consecutive grouping of probe replicates in the data file as required for calculation.

Figure 11: Consecutive grouping of replicates

2.1 Extraction filter

One of the main ideas of MADA is to allow the handling of as many different data file types as possible, regardless which software or scanner was used for producing the data. They only have to be stored in a columnar list in which each line is representing a single spot of the array and each column is holding the corresponding values of a particular parameter, such as signal intensity, background intensity, etc.. The extraction filter provides the information where the data required for further analysis are located in the list.

For some types of data files, an extraction filter is already implemented into MADA and we would like to include more of them in the future. Therefore, you are welcome to give us some feedback about the software and the type of files you are using.

However, a missing extraction filter for your type of data does NOT mean that MADA is unable to handle your data! This chapter explains how to set up custom-made extraction filters.

The location of a data subset, which correspond to a single parameter (column), can be defined via four identifiers: Data block start, data block end, data column at block and lines of data column to use.

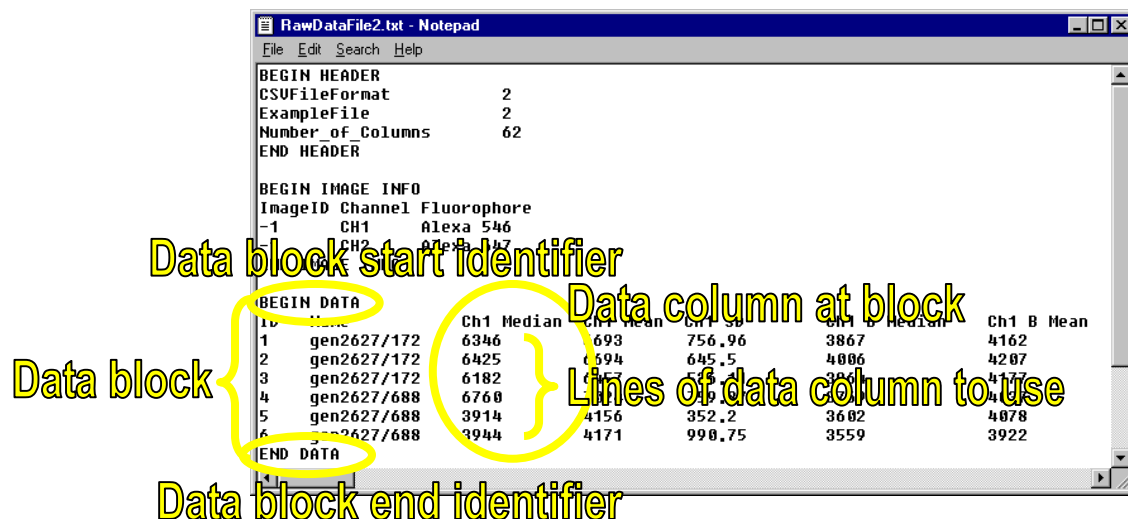


Figure 12: Identifiers for data extraction

While creating an extraction filter, instructions must be defined for every spot-describing parameter of interest using these identifiers. To create or edit an extraction filter press the 'Edit filter' button in the 'Extract data' dialog window.

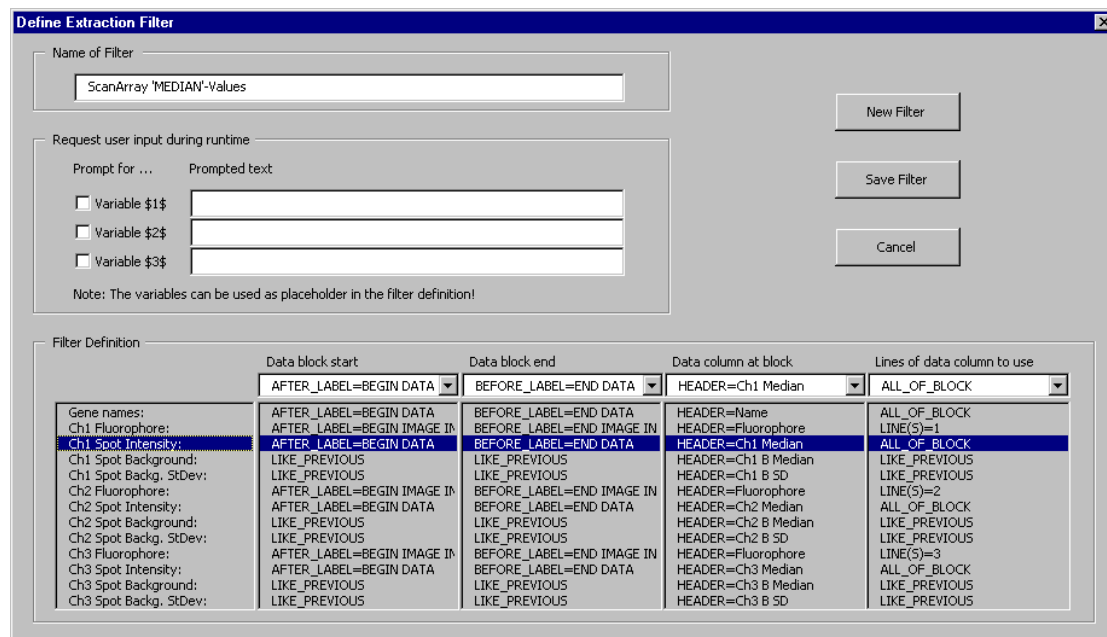


Figure 13: Setting up an extraction filter

The extraction filter currently marked will be displayed and can now be edited. To use it as a template for a new filter, change the 'Name of Filter' before saving. If you prefer to start with a blank formula, press 'New Filter'. 'Save Filter' will save your settings under the name listed in the MADA folder using the file extension *.mef* (mada extraction filter). Like all *mef*-files within this folder, it is directly available within the filter selection list in the 'Extract Data' dialog window.

2.1.1 Commands available for setting up an extraction filter

Data block start

- **AFTER_LABEL={label}** : The data block starts after the label indicated.
- **AT_LABEL={label}** : The data block starts directly at the label indicated. Can be used if there is no separate block start label.
- **AT_ROW={number}** : The data block starts at the line number indicated. Should only be used if the data starts always at the same position in the file. Fastest method because there is no need to search for a label.
- **LIKE_PREVIOUS** : Block start is identical with the one defined for the previous parameter. This is just to increase the speed of the extraction because it is faster than searching again for a similar label.

A block start must always be defined, except for fluorophores. Their names are often located within the data files' general header information and not included in the particular blocks.

Data block end

- `BEFORE_LABEL={label}`: The data block ends before the label indicated.
- `AT_ROW={number}` : The data block ends at the line number indicated. Should only be used if the data block location (starting row) and the number of rows is not varying between different raw data files. Alternatively the `LINE(S)` setting within the 'lines of data column to use' option can be used.
- `LIKE_PREVIOUS` : Block end is identical with the one defined in the previous line.

A block end must not be necessarily defined. It is only required if you want to use the `ALL_OF_BLOCK` setting within the 'lines of data column to use' option.

Data column at block

- `HEADER={label}` : The data column can be identified via the label indicated.
- `AT_COL={number}` : The data column can be identified via the number indicated. For different raw data files, assure that the particular data subsets are located within the same column. **Note:** This command is the only way to extract data columns without any header label.

A data column must always be defined, except for fluorophores (see above).

Line of data column to use

- `ALL_OF_BLOCK` : Use all data lines of the data block.
- `ALL_COUNTED` : Determine the number of data lines by counting them. To use if it was not possible to define a data block. The lines will be counted starting at the defined block start until the occurrence of an empty cell. An empty cell at the end of the data column is needed, but there should be no empty cell in between the data column. Wherever possible, use the faster and securer `ALL_OF_BLOCK` statement.
- `LIKE_PREVIOUS` : Use the same number of data lines as in the in previous one.

- LINE(S)={number} : Use the number of data lines indicated or for fluorophores use the data line indicated, respectively.
- NONE={text} : The line (cell) is not extracted from the file, but set to the text indicated. This can be used to assign a fluorophore name, which is not available in the raw data.
- NONE : The column will not be extracted. Should be used if the parameter is not of interest. For example, if there are no values for a third channel.

There must be a definition for 'line of data column to use' in any case.

Variables

Up to three variables can be used as placeholders. A prompt during runtime will ask the user to define the values required before starting the extraction. Select the variables to use and type the message to prompt in the list box. In the extraction filter definition use the placeholder \$1\$, \$2\$ or \$3\$. During runtime this placeholders will be replaced by the appropriate user input.

In Example 3, the use of variables is shown.

2.1.2 Example 1

The easiest way to set up an extraction filter is to define the data subsets of interest by their 'coordinates' in terms of rows and columns within the raw data file. This is shown in the example below (definition of the data column start position and three methods for the definition of the data column end position - by typing in directly the number of data lines via the LINE(S) tag, by defining the end row of a data column, and by counting the data lines).

Note: In this case it is a prerequisite that the locations of the data subsets of interest within the different data files are not changing (they must be located at 'fixed' locations) and, except for method 3, it must be assured the data files have always the same length.

	A	B	C	D	E	F	G	H	I	J
1		Channel 1			Channel 2			Channel 3		
2	Name	Spot Intensity	Backgr. Intensity	Backgr. StDev	Spot Intensity	Backgr. Intensity	Backgr. StDev	Spot Intensity	Backgr. Intensity	Backgr. StDev
3	gen2627/172	6346	386	20,66	3655	24	9,12	3595	31	7,13
4	gen2627/172	6425	400	20,2	3684	24	9,46	3620	30	9,1
5	gen2627/172	6182	396	20,1	3625	22	9,73	3562	29	6,23
6	gen2627/172	6283	401	22,11	3730	22	10,25	3645	27	11,15
7	gen2627/172	6450	403	21,15	3604	24	9,02	3492	26	7,02
8	gen2627/688	6760	376	20,25	3576	29	9,48	3526	27	9,46
9	gen2627/688	3914	360	20,54	3629	31	8,33	3584	34	8,23
10	gen2627/688	3944	355	18,79	3511	38	8,53	3463	19	4,57
11	gen2627/688	3658	364	17,99	3352	18	8,21	3319	25	7,26
12	gen2627/688	3616	388	20,42	3543	26	8,1	3474	23	9,14

Figure 14: Data file organization

Define Extraction Filter

Name of Filter:

New Filter

Request user input during runtime

Prompt for ...	Prompted text
<input type="checkbox"/> Variable \$1\$	<input type="text"/>
<input type="checkbox"/> Variable \$2\$	<input type="text"/>
<input type="checkbox"/> Variable \$3\$	<input type="text"/>

Note: The variables can be used as placeholder in the filter definition!

Save Filter

Cancel

Filter Definition

	Data block start	Data block end	Data column at block	Lines of data column to use
Method 1	Gene names: Ch1 Fluorophore: Ch1 Spot Intensity: Ch1 Spot Background: Ch1 Spot Backg. StDev:	AT_ROW=2 AT_ROW=2 LIKE_PREVIOUS LIKE_PREVIOUS	AT_COL=1 AT_COL=2 AT_COL=3 AT_COL=4	LINE(S)=10 NONE=Cy3 LINE(S)=10 LIKE_PREVIOUS LIKE_PREVIOUS NONE=Cy5
Method 2	Ch2 Fluorophore: Ch2 Spot Intensity: Ch2 Spot Background: Ch2 Spot Backg. StDev:	AT_ROW=2 LIKE_PREVIOUS LIKE_PREVIOUS	AT_ROW=12 LIKE_PREVIOUS LIKE_PREVIOUS	ALL_OF_BLOCK LIKE_PREVIOUS LIKE_PREVIOUS LIKE_PREVIOUS
Method 3	Ch3 Fluorophore: Ch3 Spot Intensity: Ch3 Spot Background: Ch3 Spot Backg. StDev:	AT_ROW=2 LIKE_PREVIOUS LIKE_PREVIOUS	AT_COL=8 AT_COL=9 AT_COL=10	NONE=Alexa ALL_COUNTED LIKE_PREVIOUS LIKE_PREVIOUS

Figure 15: Three extraction filter methods for a data file with 'fixed' positions for the data subsets of interest

2.1.3 Example 2, Filter for ScanArray file

Data files produced by the PerkinElmer ScanArray software are containing several 'blocks' such as a header block, an image info block and the data block. Each block is beginning with a block start tag and terminated by a block end tag. The number of data lines in a block can differ depending on the method applied during former image analysis, the number of channels available, or the number of spots analysed. Also the order and number of data columns in a block can differ depending on the settings in the image analyses software. In the following example for an extraction filter, the block start tag is used to find the first data line and the block end tag to calculate the number of data lines. The columns are identified via their headers.

	A	B	C	D	E	F	H	I	J	K	L	M	N	O	P
1	BEGIN HEADER														
2	PerkinElmer Inc.														
3	ScanArrayCS	2													
4	ScanArray Ex	2													
5	Number of C	62													
6	END HEADER														
7															
8	BEGIN GENERAL INFO														
9	DateTime	04.02.2004 16:29													
10	GalFile														
11	Scanner	Model: Express													
12	User Name														
55	BEGIN IMAGE INFO														
56	ImageID	Channel	Image	Fluorophore	Barcode	Units	Y Units	FX Offset	Y Offe	Status					
57	-1	CH1	D:\c1.r	Alexa 546		µm	10	0	0	Control Image					
58	-1	CH2	D:\c1.r	Alexa 647		µm	10	0	0						
59	END IMAGE INFO														
60															
61	BEGIN DATA														
62	Index	Name	ID	X	Y	Diameter	B Pixels	Footprint	Flags	Ch1 Median	Ch1 Mean	Ch1 SD	Ch1 B Median	Ch1 B Mean	Ch1 B SD
63	1	gen2627/172		8571	13267	274	800	12	1	6346	6693	756,96	3667	4162	2043,66
64	2	gen2627/172		9091	13276	260	940	21	1	6425	6694	645,5	4006	4207	2000,2
65	3	gen2627/172		9581	13272	240	940	13	1	6182	6457	536,11	3964	4177	2070,1
66	4	gen2627/172		10063	13277	400	364	6	1	6283	6513	768,73	4016	4327	2276,11
67	5	gen2627/172		10553	13287	400	364	17	1	6450	6729	702,5	4034	4347	2117,15
68	6	gen2627/688		11043	13277	400	364	26	1	6760	7024	759,87	3769	4037	2032,25
69	7	gen2627/688		11553	13287	400	364	17	1	3914	4156	352,2	3602	4078	2035,54
70	8	gen2627/688		12043	13287	400	364	27	1	3944	4171	990,75	3559	3922	1886,79
71	9	gen2627/688		12543	13277	400	364	26	1	3658	3949	156,05	3640	3844	1747,99
72	10	gen2627/688		13068	13282	260	940	2	3	3616	3921	123,61	3682	4223	2069,42
73	END DATA														

Figure 17: ScanArray file structure

Figure 16: Extraction filter for ScanArray files

2.1.4 Example 3, Filter for GenePix file

The GenePix software of Axon Instruments creates data files, which have first some lines of header information and starting after that directly with the data. Because the number of header lines differs to the software version and block identifiers are not used, the extraction strategy is to use the header of the first column to find the block start and count how many data lines are available. The columns are identified via the header, but because GenePix is using the wavelength as part of the headers, variables are used to prompt the user for the wavelength he is currently using to determine between the different channels. The prompted variables give the flexibility to use this filter for every GenePix file no matter which flourophores are currently used.

Block	Column	Row	Name	ID	X	Y	Dia.	F635 Median	F635 Mean	F635 SD	F635 CV	B635	B635 Median	B635 Mean	B635 SD
33	1	1	TFC3	YAL001C	2910	26140	100	138	450	824	183	0	0	1	2
34	1	2	YAL018C	YAL018C	3080	26140	70	53	134	179	133	0	0	4	20
35	1	3	FUN19	YAL034C	3270	26140	90	261	724	871	120	0	0	1	4
36	1	4	SPC72	YAL047C	3440	26130	120	240	461	476	103	0	0	1	2
37	1	5	YAL065C	YAL065C	3610	26120	110	978	978	503	51	0	0	2	3
38	1	6	CDC15	YAR019C	3780	26130	100	87	280	401	143	0	0	2	3
39	1	7	EFB1	YAL003W	3950	26130	120	14877	11388	5421	47	0	0	2	3
40	1	8	ATS1	YAL020C	4120	26120	110	1205	1121	271	24	0	0	2	4
41	1	9	YAL035C	YAL035C	4290	26120	120	351	309	121	39	0	0	1	3
42	1	10	YAL049C	YAL049C	4470	26120	120	4796	4134	1607	38	0	0	1	2
43	1	11	YAL066W	YAL066W	4630	26110	120	319	285	108	37	0	0	2	3
44	1	12	YAR023C	YAR023C	4810	26100	120	442	404	131	32	0	0	2	3

Figure 19: GenePix file structure

Figure 18: (Exemplary) Extraction filter for GenePix files

This page is intentionally left blank to support double-sided printouts.

3 Calculate

This chapter guides you through the mathematical and statistical data processing procedure offered by MADA and gives you an overview on the methods available.

Data analysis in MADA encompasses the following aspects:

- 'Local background correction' for each spot.
- 'Signal significance test' for each spot.
- 'Outlier test' to identify outliers within probe replicates.
- Calculation of 'Mean of replicates' and data normalisation.
- Calculation of 'Ratio & Intensity' including lowess curve smooth normalisation option for further analysis.

a) Pressing 'Calculate' in the MADA main window will open the 'Select data for calculation' window. Select the worksheets to calculate. If more than one worksheet is selected, it can be chosen between the modes 'Calculate all selected sheets the same way' or 'Ask at every sheet for calculation steps to do'. Press 'NEXT' to continue.

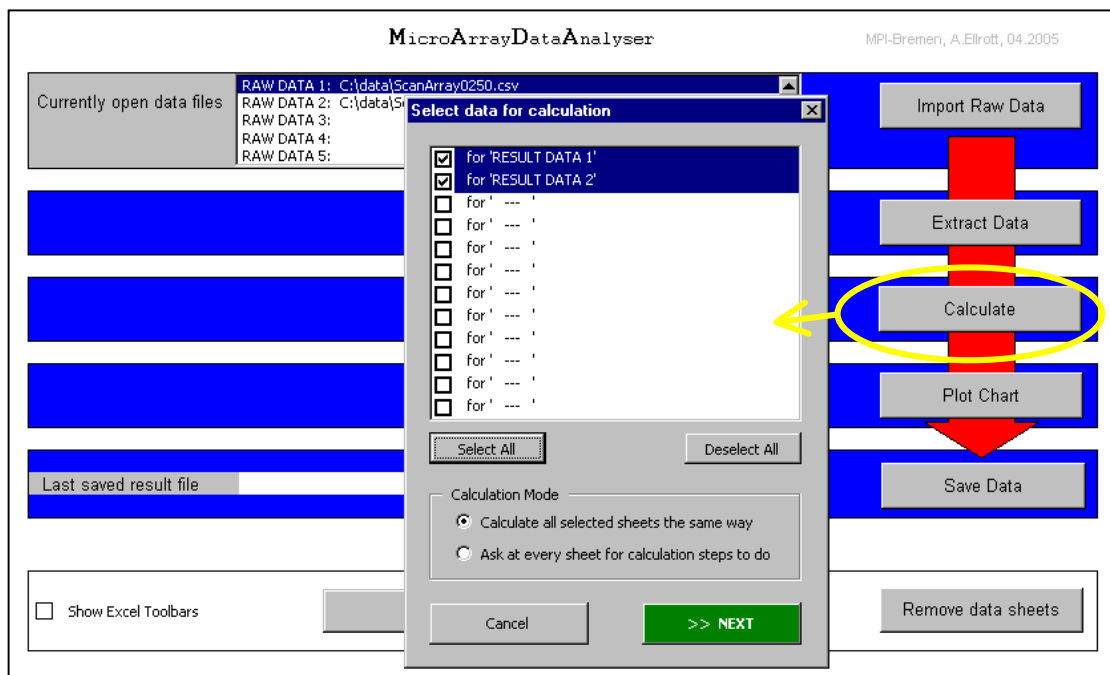


Figure 20: Selection of the data sets to process

b) In the 'Select calculation steps' window the workflow of MADA's data processing is visualised and settings can be adjusted.

- Set the 'Minimum number of replicates that must be assigned as signal'.
It is recommended to use: $(\text{number of replicates} / 2) + 1$
- Select a normalisation approach (optionally).
- Set the channels to calculate 'Ratio & Intensity' (only if more than one channel is available).

Figure 21: Select calculation steps

The button 'Advanced options' enables access to more detailed settings.

Settings of the significance and outlier test can be changed and calculation methods can be ignored or whole calculation steps can be disabled.

Note: There is a clear difference between ignoring a calculation method and disabling the calculation step!

- Ignoring a calculation method means that only a predefined output value is created to prevent errors during further calculation steps.
- Disabling a calculation method means that there is no corresponding calculation at all. However, it is quite likely that the following calculation step is reporting an error because of missing values. A step should only be disabled, if the value is not needed for further calculation or if it was already calculated, previously. It is an option for speeding up recalculations or to prevent overwriting of manually modified results.

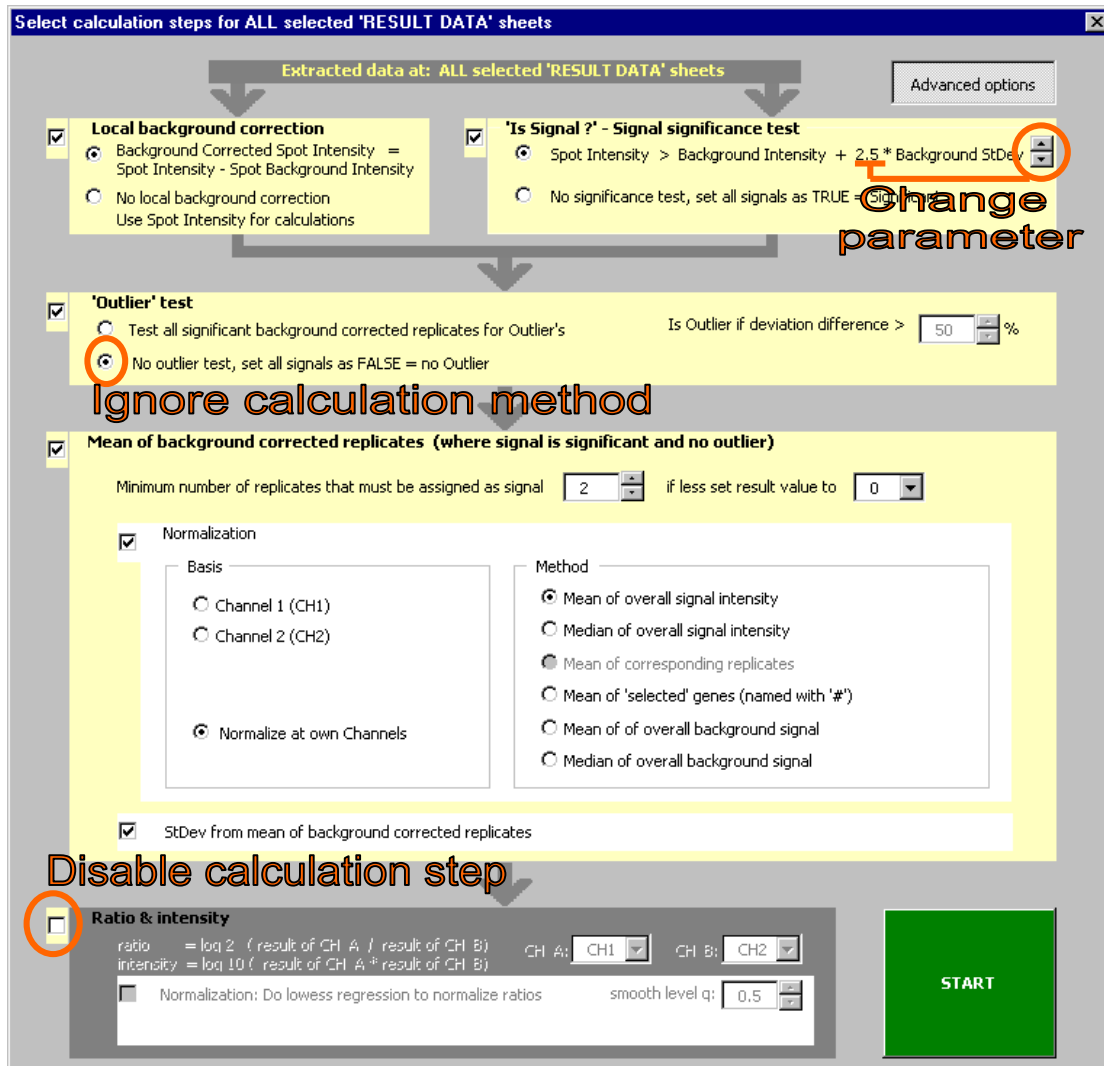


Figure 22: Advanced options of the 'Calculation' menu

- c) Press 'Start' to begin the calculation. A progress bar will inform you about the current status of data processing.

Depending on the size of the data set and the performance of your system, calculations can take several minutes. Refer to *System requirements and Performance* for additional information.

Depending on the calculation mode selected, MADA will prompt for new settings for every worksheet.

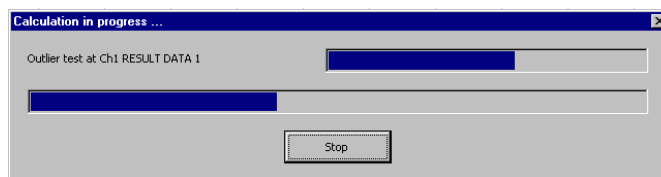


Figure 23: Progress of calculation

3.1 Local background correction

Labeled target molecules do not only hybridise with the surface-immobilised capture probes, normally, they also bind to the microarray surface itself, resulting in clear background signals. Therefore, it must be assumed that part of each “hybridisation” signal is also based on local background fluorescence. This part has to be subtracted from the total signal, but the fraction of the background signal underlying a spot signal can only be estimated.

Normally, the local background directly surrounding a spot should represent the fraction of the spot signal which is caused by adsorption to the slide surface:

$$\text{Background Corrected Spot Intensity} = \text{Spot Signal Intensity} - \text{Spot Background Intensity}$$

The local background correction can be disabled by selecting 'No local background correction' within the 'Advanced options'.

3.2 Signal significance test

To make sure that a spot signal is really a signal and not just noise, a statistical test is required. A test for significance based on a one-sided t-test reduces the probability of getting false positive signals.

$$\text{spot signal intensity} > \text{local background intensity} + t * \text{local background standard deviation}$$

The t-score (t) that corresponds to a given probability depends on the degree of freedom (df) and can be found in Table 1. The df-value correlates to the number of image pixels which are used to determine the local background minus one. Since the t-distribution converges to the normal distribution at high values of df, the t-score can be approximated by the z-value if $df \geq 30$ [Bortz]. Usually, the raw data is based on high-resolution microarray images which allows an approximation via the z-value in most cases. The z-value of the standard normal distribution is equal to the t-score for $df = \infty$ (Table 1).

Probability that the spot intensity is significantly higher than the local background intensity (Probability that the spot intensity is a false-positive signal)									
df	70% (30%)	80% (20%)	85% (15%)	90% (10%)	95% (5%)	97.5% (2.5%)	99% (1%)	99.5% (0.5%)	99.9% (0.1%)
1	0.727	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.313
2	0.617	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
5	0.559	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
10	0.542	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.143
15	0.536	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
20	0.533	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
30	0.530	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385
60	0.527	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.232
120	0.526	0.845	1.041	1.289	1.658	1.980	2.358	2.617	
$\infty = z$	0.524	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090

Table 1: Critical values (t-scores) of the Student's t-distribution for given probabilities at different degrees of freedom (df). [Bortz], [NIST/SEMATECH]

In statistical analysis, the two categories 'significant' $\geq 95\%$ and 'highly significant' $\geq 99\%$ are commonly applied [Bortz]. MADA is using a default t-score of 2 which approximately corresponds to a 97.5% probability that the signal intensity is significantly higher than the local background. The t-score can be adjusted via spin buttons at the 'ADVANCED SETTINGS' menu. It is also possible to disable the significance test by selecting 'No significance test ...'.

3.3 Outlier test

A certain gene is not only represented by a single spot on the chip. Normally, three or more spot replicates of a single capture probe are printed. In a 'perfect' experiment, all replicates of a particular probe will show identical signal intensities without any deviation. In reality, replicate spots often possess clear deviations in their intensity. For example, dirt on the slide surface can lead to a high local fluorescence (see picture, replicate10) and artifacts, based on the spotting procedure, to low signal intensities (replicate 4). Most image analysis softwares allow to manually remove these spots from the data set. However, such a process is completely subjective as well as extremely time consuming (not practical for chips carrying several thousands of probes).



Figure 24: Outliers within replicates

Our idea was to automate this process. If you have a look on the 10 replicates in the image exemplary shown, it is obvious that spots 4 + 10 represent 'bad' spots (outliers) because they are quite different from the rest. The standard deviation from the signal intensities of this ten replicated spots can now be calculated. Because of the two outliers it will result in a comparable high value. What happens if we remove one of the data points and calculate the standard deviation again? If we remove one of the 'nice' spots, the value will not change remarkably but if we calculate it without one of the outlying spots, the deviation value will dramatically decrease: an outlier is found! This is the general idea of the outlier test algorithm implemented in MADA.

By default, settings the minimum deviation difference to declare a spot as an outlier is set to 50%. It can be adjusted in the 'ADVANCED OPTIONS' menu. If no outlier test is desired, select 'No outlier test ...'.

Note: To perform this test, the array must carry at least triplicates for each probe since a minimum of three significant spot signals is required.

```

devAll = deviation over all replicates
devPossibleOutlier = devAll
do while idOutliersToTest < (numOfReplicates / 2) + 1
  idOfPossibleOutlier = 0
  for idTestReplicate = 1 to numOfReplicates
    devTest = deviation over all replicates, except replicate(idTestReplicate), except found outliers
    if devTest is defined % < devAll and devTest < devPossibleOutlier then
      idOfPossibleOutlier = idTestReplicate
      devPossibleOutlier = devTest
    end if
  next idTestReplicate
  if idOfPossibleOutlier > 0 then
    replicate(idOfPossibleOutlier) = found outlier
    devAll = devPossibleOutlier
  else
    exit
  end if
loop
    
```

Figure 25: Basic principle of the outlier test algorithm

3.4 Mean of Replicates

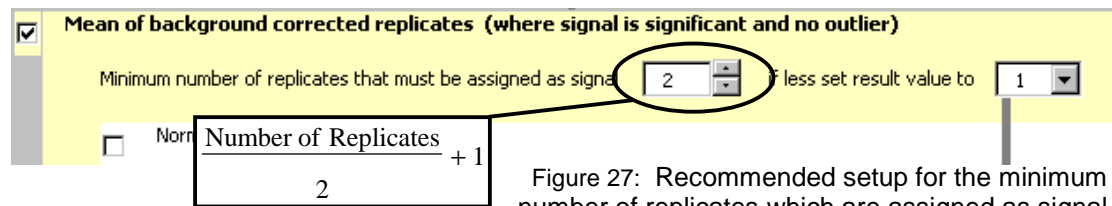
In this step, the replicates of each probe will be combined to a single value by calculating their arithmetic mean. Only significant and non-outlying replicates will be considered.

The mean of replicates can only be calculated if there is at least the number of valid signals specified.

Name	Alexa 647							
	Backg. Intensity	Is Signal?	Has Outlier?	Mean of Rep. of Backg. Corr. Intensity	StDev of Rep. of Backg. Corr. Intensity	Backg. Intensity	Backg. Intensity	Backg. Intensity
GeneA	115.77	11.81	53.82	103.96	0	0		
GeneA	122.74	26.48	78.13	96.26	0	0		
GeneA	335.36	4.11	32.26	331.24	1	0		0
GeneB	1611.19	33.45	119.46	1577.74	1	0		
GeneB	1670.26	52.63	210.64	1617.63	1	0		
GeneB	1299.20	2.35	24.47	1296.85	1	1	1597.69	28.20615433
GeneC	1406.84	26.76	87.27	1380.07	1	0		
GeneC	1359.66	6.97	41.75	1352.69	1	0		
GeneC	1376.65	14.08	66.76	1362.57	1	0	1365.11	13.86721776

Figure 26: Calculation of 'Mean of Replicates', example for triplicated probes

This number can be adjusted via the spin buttons. To obtain reasonable results, at least the half of the replicates plus one should have a valid signal. It is recommended to setup the 'Minimum number of replicates that must be assigned as signal' to:



If the calculation for a certain replicate fails, the corresponding mean of replicate value will be set to zero by default. Differential gene expression analysis is based on the calculation of differences between two states, so normally significant signals of both states / channels are required to do a calculation. However, it is possible that a certain gene is expressed under condition B, but not at all under condition A. Since a value of zero can not be processed, these events are normally “hidden” during later analysis. As a workaround, a certain value can be assigned in such cases, e.g., ‘1’ as shown above. NOTE: If you use this option, you potentially will get hints for additional regulation events but you have no information on the numeric expression level (ratio) for these data points. Be aware that you might also get a number of false positive regulation events.

By default, the standard deviation of the valid replicates is automatically calculated, but it can be disabled in the 'Advanced options' mode.

3.4.1 Normalisation

If desired, data normalisation can be performed by activating the checkbox and choosing a method. Depending on your choice, MADA will calculate a particular normalisation factor which will adjust all 'mean of replicates' values based on a certain standard.

3.4.1.1 Mean / Median of overall signal

The normalisation factor is calculated by the arithmetic mean or median of the background-corrected signal intensities of all significant and non-outlying spot signals.

In case of an array with a high number of spots (e.g. whole genome arrays), the differences between single spots will be compensated by the size of the data set, so that the mean value of all spots can be used as a standard.

3.4.1.2 Mean of corresponding replicates

Normalisation based on the 'Mean of corresponding replicates' means that each mean of replicate value is directly normalised to its corresponding mean of replicate value of the selected channel.

Example: Channel 2 is selected as the normalisation standard. Mean of replicate <probe A> of channel 1 will be normalised to mean of replicate <probe A> of channel 2. Mean of replicate <probe B> of channel 1 will be normalised to mean of replicate <probe B> of channel 2, etc.

At least two fluorescent channels must be available to use this method. It is disabled if 'Normalise at own channel' is selected.

3.4.1.3 Mean of selected genes

Often, so called 'house keeping genes' are used as a reference for normalisation. If you define the signal of selected genes for the normalisation procedure, they should be assumed as not regulated.

To define genes, their (spot) name has to start with the hash (#) character in the 'RESULT DATA' worksheet.

Pressing the '>> SELECT'-button will open a gene selection menu. First, select the data worksheet at the top bar, then select the appropriate genes in the list. 'Save changes' will automatically insert a leading hash character to the selected gene names within the result data set.

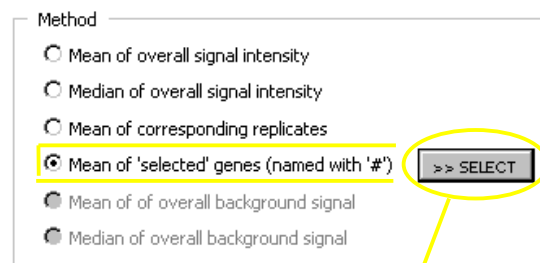


Figure 28: Mean of selected genes

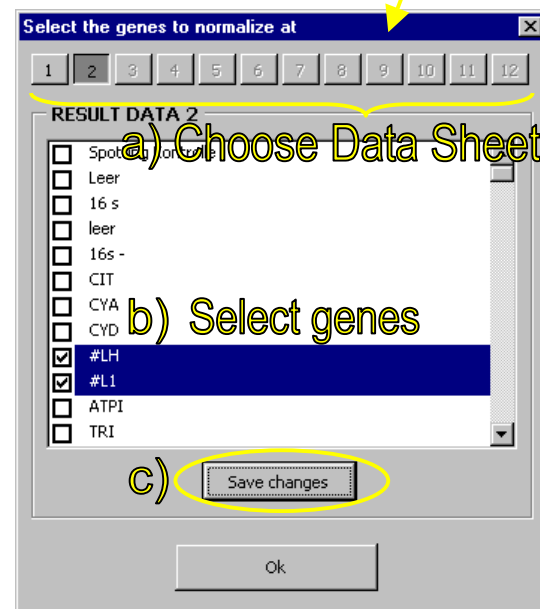


Figure 29: Gene selection menu

3.4.1.4 Mean / Median of overall background signal

The normalisation factor is calculated by the arithmetic mean or median of the local background intensities of all spots.

The method can only be chosen if 'Normalise at own channels' is selected.

3.4.1.5 Lowess normalisation

Lowess normalisation is done by applying a lowess curve smooth function to the set of ratio and intensity values. Therefore, the method can be found under the ratio & intensity calculation option and is only applicable for two channel data such as from gene expression profiling.

If you select the lowess normalisation, you should disable the other normalisation approaches of the previous calculation step, except you intent to apply a two-stage normalisation approach with one of the other normalisation method in the beginning and lowess additionally.

Lowess normalisation adjusts each data point depending on the positions of the surrounding data points. How many surrounding data points are taken into account for this calculation is selectable via the 'smooth factor'. The higher the smooth factor, the larger is the subset of data taken. For microarray normalisation a smooth factor within the range of 0.25 to 0.5 is common.

If you assign other values than zero to data points which failed the criteria for the 'mean of replicates' calculation (Chapter 3.4), it is recommended to not use this biased values for the lowess regression calculation since they could introduce significant artificial shifts.

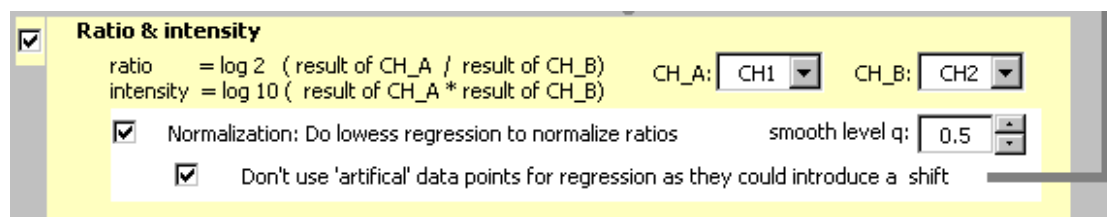


Figure 30: Lowess normalisation

On the next page, an example is given to explain how lowess regression calculation works in detail and which calculation steps are done to adjust a data point depending on the surrounding data points.

A Take the Intensity and Ratio values as full data set sorted by Intensities. Calculate the frame size according to the user selected smooth factor (0.33 in the example) and adjust the frame around the currently to process data point, the point of estimation, in order to get a local subset of data.

No	Intensity	Ratio	dist_min = point_of_est - Intensity	dist_max = Intensity - point_of_est	dist_diff = dist_min - dist_max	dist_frame = Int_max - Int_min	dist_sum = dist_diff + dist_frame
1	0.55782	18.63654					
2	2.021727	103.49646					
3	2.577325	150.35391					
4	3.414029	190.51031					
5	4.301408	208.70115	3.832179		3.832179	4.7195586	8.5517376
6	4.744839	213.71135	3.388748		2.3998975	4.8210295	7.220927
7	5.107378	228.49353	3.0262093		0.7698696	7.1848269	7.9546965
8	6.541166	233.55387	1.5924212		2.6537588	7.2723893	9.9261481
9	6.721618	234.55054	1.4119698		0.9888505	3.7273047	10.4590004
10	7.260058	223.89225	0.8735291		3.7960789	5.2696288	7.5551277
11	8.133587	227.68339	0	0	4.24618	7.2395152	8.1130443
12	9.122438	223.91982			5.1392745		15.3525595
13	11.92967	168.01999			6.1431579		
14	12.37977	164.9575			7.2395152		
15	13.27286	152.61107					
16	14.27675	160.78742					
17	15.3731	168.55567					
18	15.64766	152.42658					
19	18.56054	221.70702					
20	18.58664	222.6904					

Point of estimation

Adjust the frame around the point of estimation to that position there the distance sum

frame_size = smooth_factor * data_set_length (rounded to the next odd integer value)
0.33 * 21 = 6.93 -> 7

B a) Determine the distance from each point to the point of estimation, scale the distances by the maximum distance over all points in the local set, compute the weights by evaluating the tricube weight function.
 b) Perform a weighted least squares fit of the local model (a line in our case) to the data subset and evaluate the regression value at the point of estimation.

Tricube weight function
 $w = 1 - |scaled_distance|^3$ for $|scaled_distance| < 1$
 0 for $|scaled_distance| \geq 1$

Local subset of data according to the point of estimation			distance = Intensity - p_of_est.	scaled_dist = distance / max dist.	w (weights)	Int * w	w_Int° = Int - w_Int	Ratio * w	w_Ratio° = Ratio - w_Ratio	w_Int° * w_Ratio° * w	(w_Int°) ² * w
No	Intensity	Ratio									
6	4.744839	213.71135	3.3887480	1	0.000000	0.0000	-2.877715	0.0000	-14.569011	0.000000	0.000000
7	5.107378	228.49353	3.0262093	0.8930169	0.023847	0.1218	-2.515176	5.4490	0.213169	-0.012786	0.150862
8	6.541166	233.55387	1.5924212	0.4699143	0.719886	4.7089	-1.081388	168.1322	5.273509	-4.105302	0.841835
9	6.721618	234.55054	1.4119698	0.4166642	0.798309	5.3659	-0.900936	187.2439	6.270179	-4.509673	0.647976
10	7.260058	223.89225	0.8735291	0.2577734	0.949490	6.8934	-0.362496	212.5835	-4.388111	1.510328	0.124766
11	8.133587	227.68339	0	0	1.000000	8.1336	0.511033	227.6834	-0.596971	-0.305072	0.261155
12	9.122438	223.91982	0.9888505	0.2918041	0.927296	8.4592	1.499884	207.6399	-4.360541	-6.064798	2.086093
			Σ 4.4189		Σ 33.7			Σ 1008.73		Σ -13.487	Σ 4.1127

$w_Int = \sum (Int_i * w_i) / \sum (w_i)$	7.622554
$w_Ratio = \sum (Ratio_i * w_i) / \sum (w_i)$	228.2804

slope = $\sum (w_Int_i^\circ * w_Ratio_i^\circ * w_i) / \sum (w_Int_i^\circ{}^2 * w_i) = -3.2794$
 intercept = $w_Ratio - (slope * w_Int) = 253.2781$
 regression_function_value = $(slope * point_of_estimation) + intercept = 226.6045$
 normalized Ratio = $Ratio - regression_function_value = 1.07889$

The example data is no microarray data it was taken and modified from the NIST/SEMATECH e-Handbook of Statistical Methods, 'Example of LOESS Computations' at <http://www.itl.nist.gov/div898/handbook/pmd/section1/dep/dep144.htm>.

3.5 Ratio & intensity

'Ratio & intensity' is applied in microarray-mediated gene expression profiling which is normally based on two-colour hybridisations and in which the data from the two channels represent the expression profile of a single cell line/culture at two different 'conditions'. The expression ratio can be calculated as (spot signal channel A / spot signal channel B) and the intensity as (spot signal channel A * spot signal channel B).

The most widely used alternative transformation of the ratio is the logarithm base 2, which has the advantage of producing a continuous spectrum of values and treating up- and downregulated genes in a similar fashion.[Quackenbush 2002]

Example:

- Upregulation by factor 2 \rightarrow expression ratio = 2 \rightarrow $\log_2(2) = 1$
- Downregulation by factor 2 \rightarrow expression ratio = 0.5 \rightarrow $\log_2(0.5) = -1$

In MADA the logarithmic ratio and intensity is used and calculated from mean of replicate values by:

- ratio = $\log_2 (A / B)$
- intensity = $\log_{10} (A * B)$

* A, B = (normalised) mean of replicates of channel 1, 2 or 3

The channels can be set in the 'Calculation' menu. If no expression analyses should be done, the calculation can be disabled in the 'Advanced Options' menu.

Note: Ratio and intensity must be calculated if RI- or RR-Plots should be used later.

Zero or division by zero can occur if the mean of replicate value of one channel contains too much non-significant or outlying spot signal intensities.

Ratio and intensity is automatically disabled if less than two channels are available. *This page is intentionally left blank to support double-sided printouts.*

4 Plot Chart

A selection of charts and graphs for the visualisation of microarray data is directly implemented in MADA. Every chart can be adapted to own preferences by using the tools of Excel.

Data that are required for building charts or plots are copied to the 'CHART' worksheet. They can be found behind the plot or chart area.

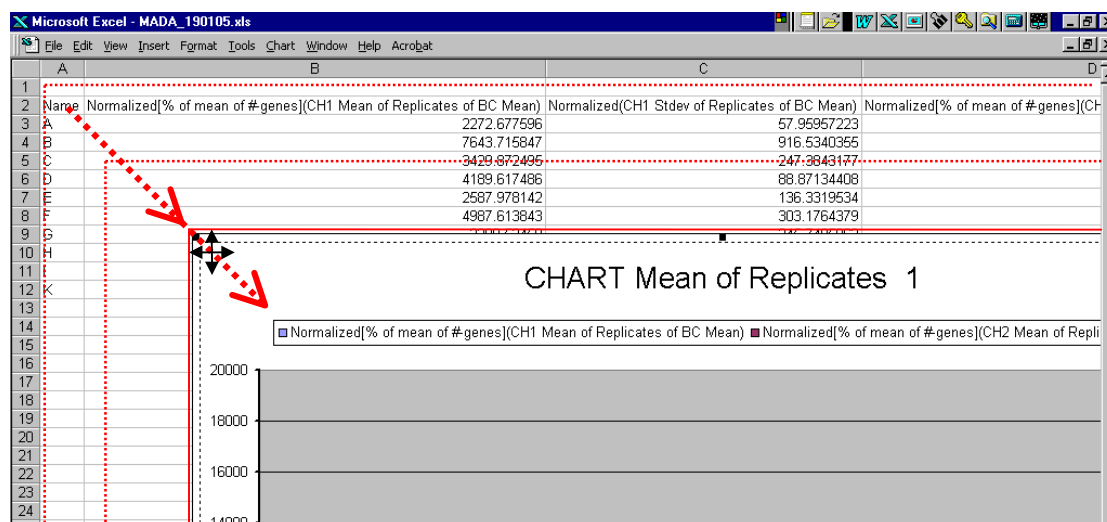


Figure 31: Finding the data which are underlying a chart or graph

Note: To save a chart or plot it is recommended to save the 'CHART' worksheet because it also includes the underlying data. Use the 'Save Data' option within the 'COMMAND' worksheet (MADA main window).

4.1 Mean of replicates

Absolute or normalised signal intensities of each probe (mean of replicates) can be visualised in a column/bar chart.

a) Press 'Plot Chart' in the MADA main window of the 'COMMAND' worksheet and select 'Plot Column-Chart: Mean of replicates' within the 'Select charts to plot' window and press '>>NEXT' to continue.

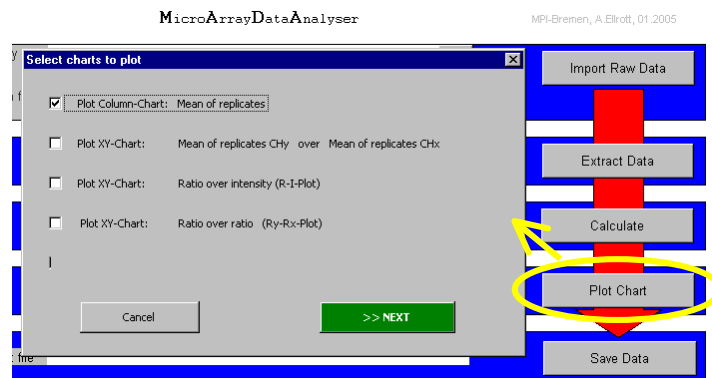


Figure 32: Selection of kind of visualization, here 'Mean of replicates'

b) Select a 'RESULT DATA' worksheet and continue by pressing '>>NEXT'.

Note: If a mean of replicate chart was already build from the same data set, it will be overwritten.

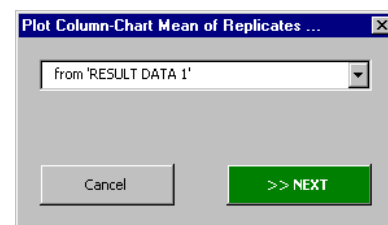


Figure 33: Selection of the underlying data set

A new worksheet for the plot is created and the underlying data are copied from the corresponding RESULT DATA worksheet.

c) Select the channels and probes which should be visualised.

Since non-significant and outlying data points have a mean of replicate value of zero, they can be automatically deselected by using the 'Select all signals > 0' option.

Press 'START' to create the column chart.

Note: In case of large data sets this may take a while.

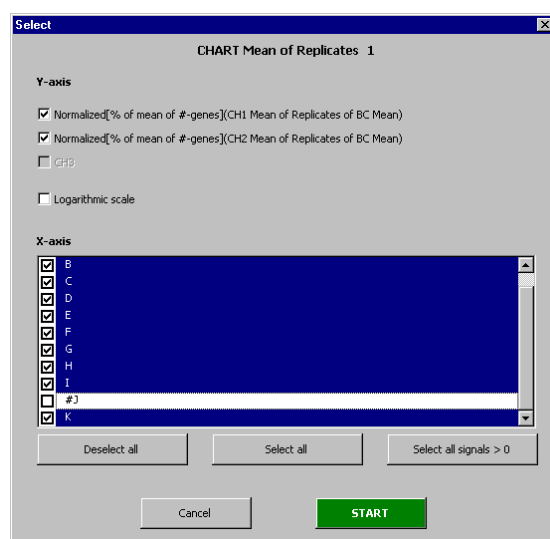


Figure 34: Selection of data points which should be visualized

The 'CHART' worksheet will be named 'CHART Mean of Replicates x' (x is the corresponding number of the data set).

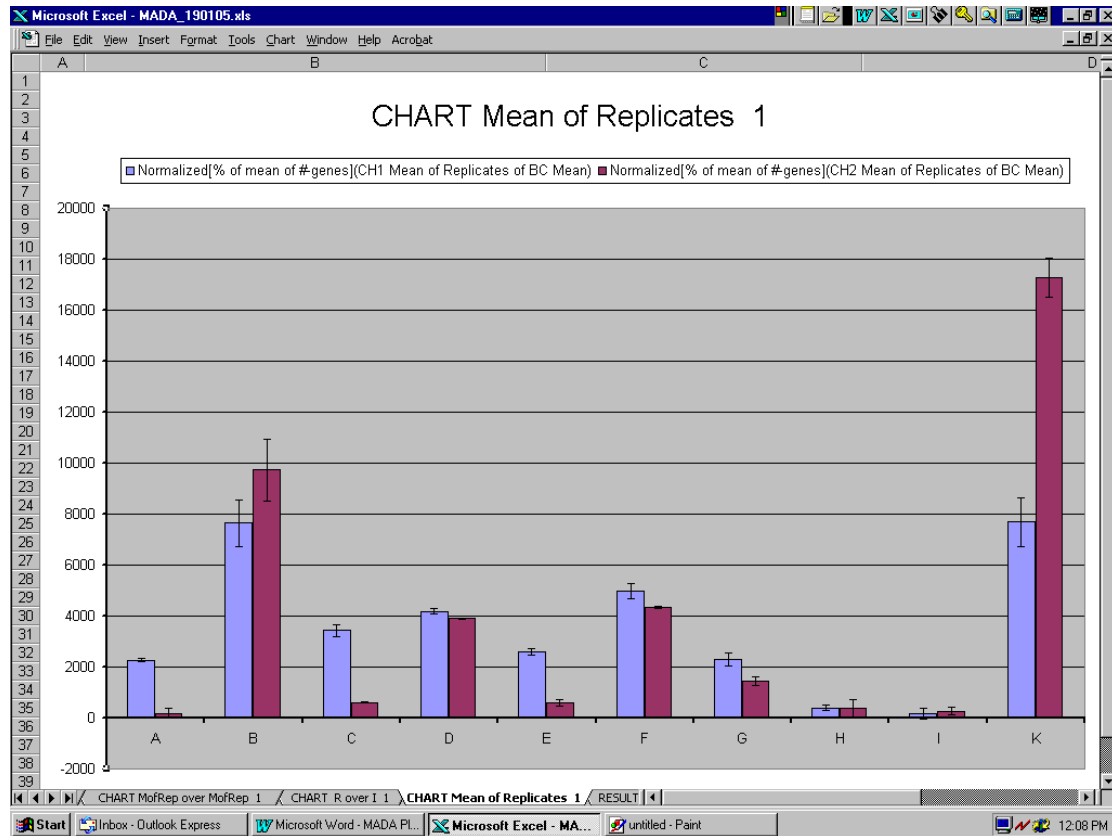


Figure 35: CHART Mean of Replicates

4.2 Mean of replicates Chy over Mean of replicates Chx

In this chart, the mean of replicate signals of channel y are plotted against the mean of replicate signals of channel x. Variations in the two data sets can easily be detected by the location of the data points compared to the bisecting line since identical values will exactly fall on that line.

a) Press 'Plot Chart' within the 'COMMAND' worksheet, select 'Plot XY-Chart: Mean of replicates CHy over Mean of replicates CHx' within the 'Select charts to plot' window, and press '>>NEXT' to continue.

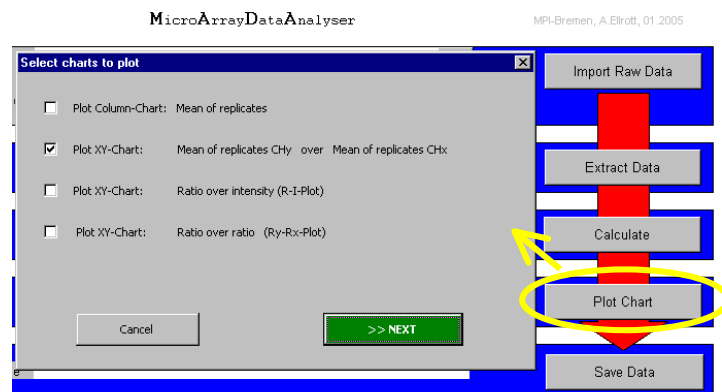


Figure 36: Selection of kind of visualization, here 'M. of Rep. CHy over M. of Rep. CHx'

b) Select worksheet to plot and continue with '>>NEXT'.

Note: If a 'Mean of Rep. over Mean of Rep.' chart generated from the same data set already exists it will be overwritten without any warning.

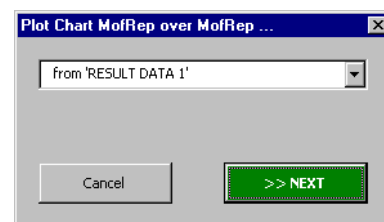


Figure 37: Selection of the underlying data set

A new worksheet for the plot is created and the data underlying are copied from the RESULT DATA worksheet selected.

c) Select the values for the x- and y-axis. Optionally, a logarithmic scaling can be applied.

Press '>> NEXT' to continue.

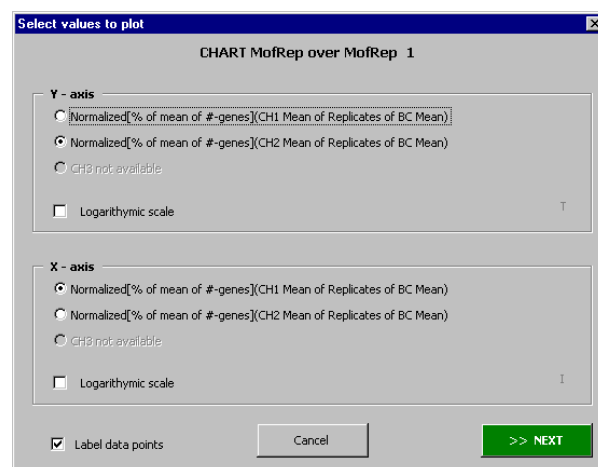


Figure 38: Selection of underlying values

d) Select the mean of replicates that should be visualised.

Since non-significant and outlying data points possess a mean of replicate value of zero, they can be automatically deselected by using the 'Select all signals > 0' option.

Press 'START' to create the XY-Chart.

Note: In case of large data sets this can take a while.

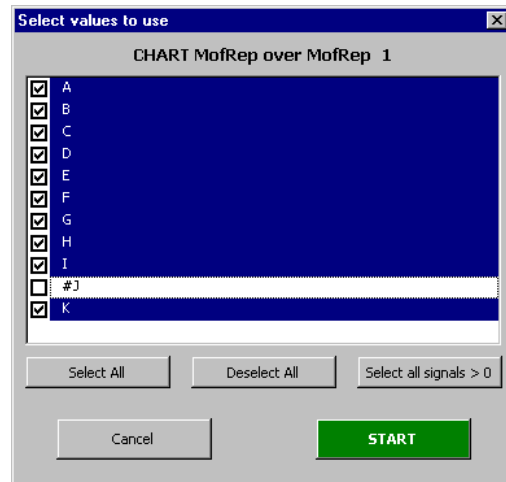


Figure 39: Selection of data points to be visualized

The 'CHART' worksheet will be named 'CHART MofRep over MofRep x' (x is the corresponding number of the data set).

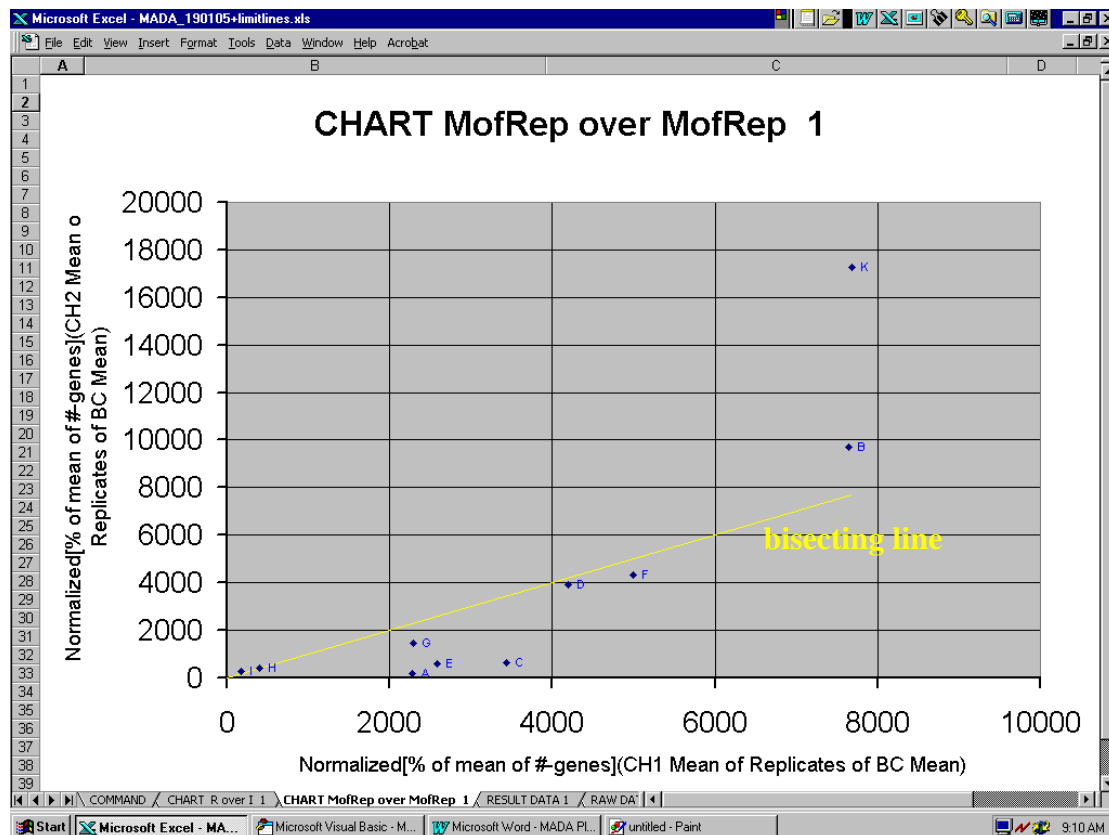


Figure 40: CHART Mean of Replicates Chy over Mean of Replicates Chx

4.3 Ratio over Intensity

R-I-plots are quite common in differential gene expression analysis since the up- and down-regulation of genes is visualised in the same manner due to a logarithm base 2-based transformation of the data. Gene regulation of a factor of, e.g., 2 is defined, for up-regulation, by a twofold increase of signal intensity (factor 2) and, for down-regulation, by a bisection of the signal intensity (factor $\frac{1}{2}$). Transferred to ratio values using the logarithm base 2 this is leading to an up-regulation of $\log_2(2) = 1$ and a down-regulation of $\log_2(\frac{1}{2}) = -1$.

Sometimes, R-I-plots are initially used in data analysis to visualise the distribution of the signals of single spots (not their replicates) and to detect, e.g., outliers. In MADA, non-significant and outlying data points will be detected during the calculation step. The R-I-plot function is implemented for visualisation of the final results and therefore it is based on ratios and intensities of mean of replicate signals and not on signal intensities of single spots.

Note: Ratio and intensity values must be calculated before plotting using the 'Calculate' function of MADA.

a) Press 'Plot Chart' in the 'COMMAND' worksheet, select 'Plot XY-Chart: Ratio over intensity (R-I-Plot)' within the 'Select charts to plot' window and press '>>NEXT' to continue.

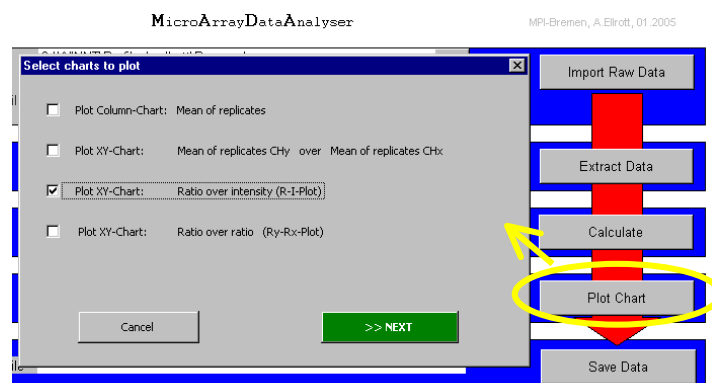


Figure 41: Selection of kind of visualization, here 'R-I-Plot'

b) Select worksheet to plot and continue with '>>NEXT'.

Note: If a 'R over I' chart generated from the same data set already exists it will be overwritten without any warning.

A new worksheet for the plot is created and the data underlying are copied from the RESULT DATA worksheet selected.

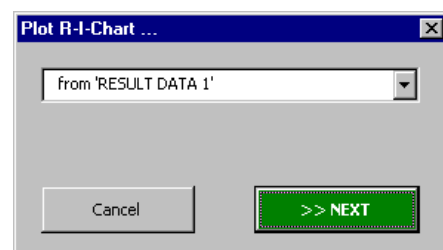


Figure 42: Selection of the underlying data set

c) Select the data points to visualise.

Note: Ratios over intensities can not be calculated from 'signals' that possess a mean of replicate value of zero. Therefore, data points can miss in the list because of a division by zero error during the previous calculation of ratios.

For expression analysis, a regulation factor, representing the threshold to indicate up- and down- regulated genes by specific colours, can be defined.

Press 'START' to create the XY-Chart.

Note: In case of large data sets this can take a while.

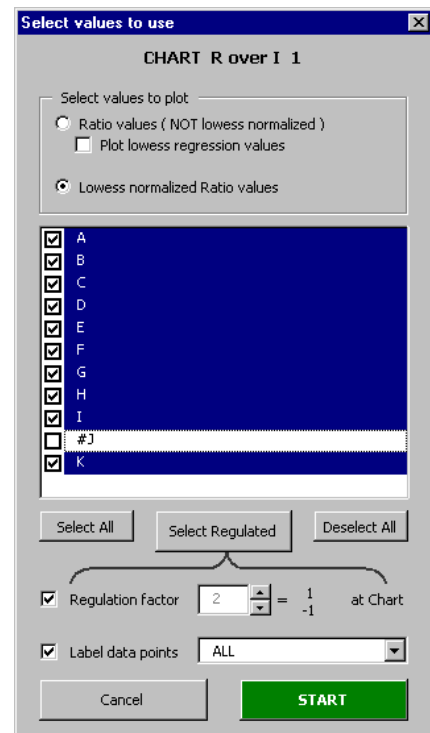


Figure 43: Selection of data points to be visualized

The 'CHART' worksheet will be named 'CHART R over I x' (where x is the corresponding number of the dataset).

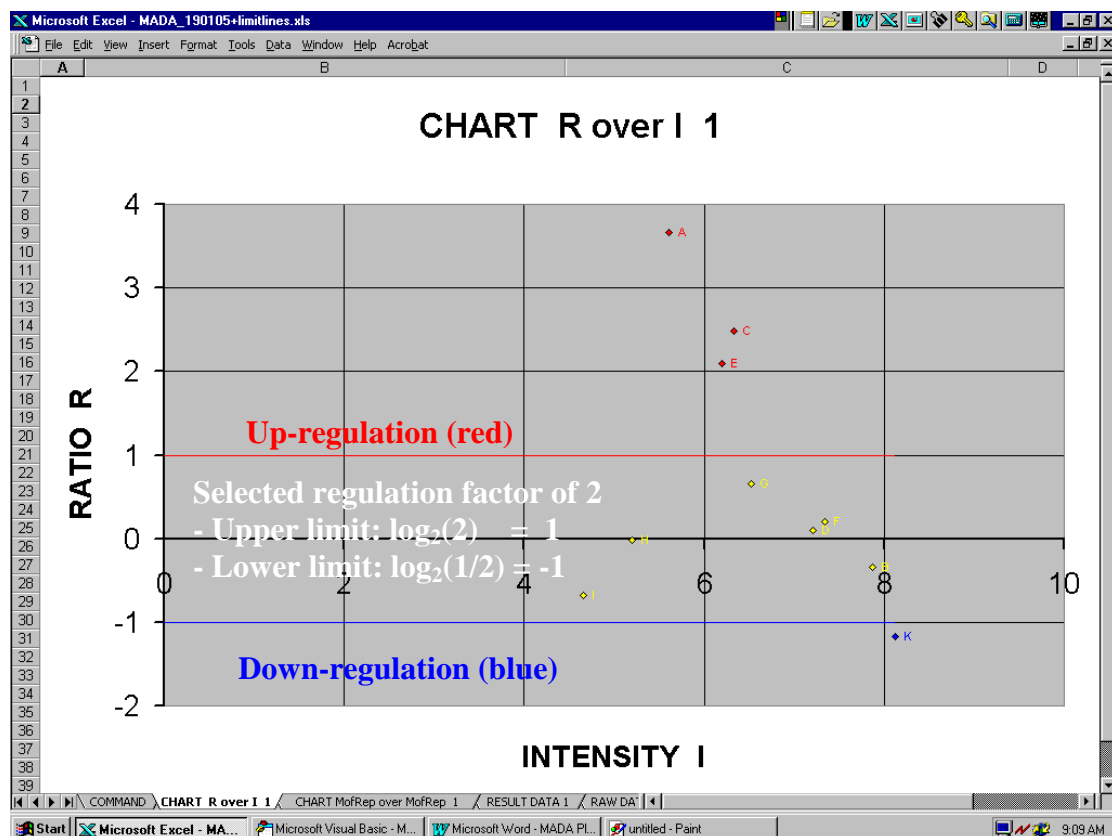


Figure 44: CHART R over I

4.4 Ratio over Ratio

This chart type is for the comparison of ratio-based results from a single array that were differently calculated or from two arrays that were differently processed.

An example for comparison of differently processed arrays is the 'flip-flop' labelling, routinely applied for two-colour hybridisation experiments. Here, the hybridisation of the two samples under investigation is repeated under identical conditions, except that the fluorescent dyes for labelling of the two pools of target molecules are exchanged. This is done to account for potential biases caused by the different dyes. Ideally, the results should be the same for both experiments, indicated by the data points which invariably fall onto the bisecting line.

Note: If the channel assigned to a particular fluorescent dye is switched in one raw data set, do not forget to switch the channel assignment for ratio and intensity calculation, too. Example:

Data set	Fluorescent dye at CH1	Fluorescent dye at CH2	Channel assignment at calculation step	Calculation
1	Cy3	Cy5	CH_A = CH1 CH_B = CH2	ratio = $\log_2(\text{CH}_A / \text{CH}_B)$ ratio = $\log_2(\text{CH1} / \text{CH2})$ = $\log_2(\text{Cy3} / \text{Cy5})$!
2	Cy5	Cy3	CH_A = CH2 CH_B = CH1	ratio = $\log_2(\text{CH2} / \text{CH1})$ = $\log_2(\text{Cy3} / \text{Cy5})$!

A single raw data set can be imported multiple times and assigned to several MADA data sets to analyse it in parallel with different calculation methods. Here, the 'Ratio over Ratio' plot should be helpful to find variations.

Data points falling next to the central bisecting line possess equal values, whereas data points falling beyond the two outer bisecting lines possess a deviation of more than 50%.

Note: It is not possible to compare values from arrays organised in a completely different way. It must be assured that the raw data sets have a comparable structure including a similar number of spots, similar spot names and the data lines ordered in a similar way. Extraction of data using MADA's 'Data Extraction' tool must be done with identical settings, especially if using 'Consecutive grouping'.

Note: Ratio and intensity values must be calculated before plotting using the 'Calculate' function of MADA.

a) Press 'Plot Chart' in the 'COMMAND' worksheet, select 'Plot XY-Chart: Ratio over ratio (Ry-Rx-Plot)' within the 'Select charts to plot' window and press '>>NEXT' to continue.

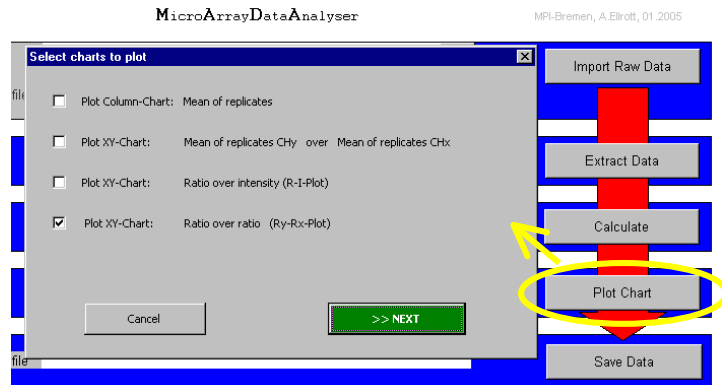


Figure 45: Selection of kind of visualization, here 'Ry-Rx-Plot'

b) Select worksheets to plot and continue with '>>NEXT'.

Note: If a 'R over R' plot generated from the same data set already exists it will be overwritten without any warning.

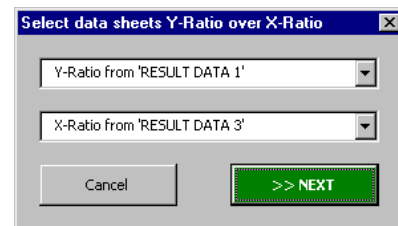


Figure 46: Selection of the underlying data set

If one of the selected worksheets contains both types of ratio values, a 'usual' and a lowess normalised one, you will be asked to select the one to plot.

Note: The comparison of data sets with different normalisation methods applied might lead to artificial shifts introduced by the differences of the normalisation methods.

A new worksheet for the plot is created and the data underlying are copied from the RESULT DATA worksheet.

c) Select the data points which should be visualised.

Note: Ratios can not be calculated from 'signals' that possess a mean of replicate value of zero. Therefore, data points can miss in the list because of a division by zero error during the previous ratio calculation.

Press 'START' to create the XY-Chart.

Note: In case of large data sets this can take a while.

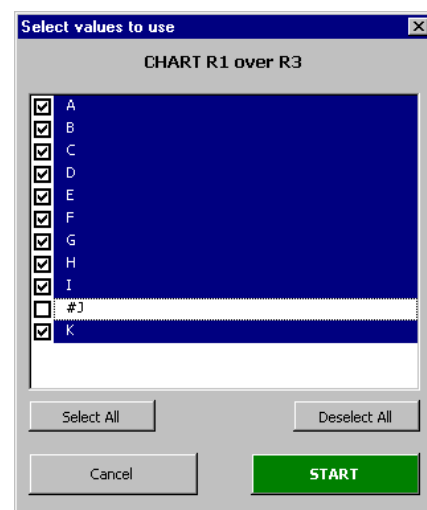


Figure 47: Selection of data points to be visualized

The 'CHART' worksheet will be named 'CHART Ry over Rx' (where x and y are the corresponding numbers of the datasheets).

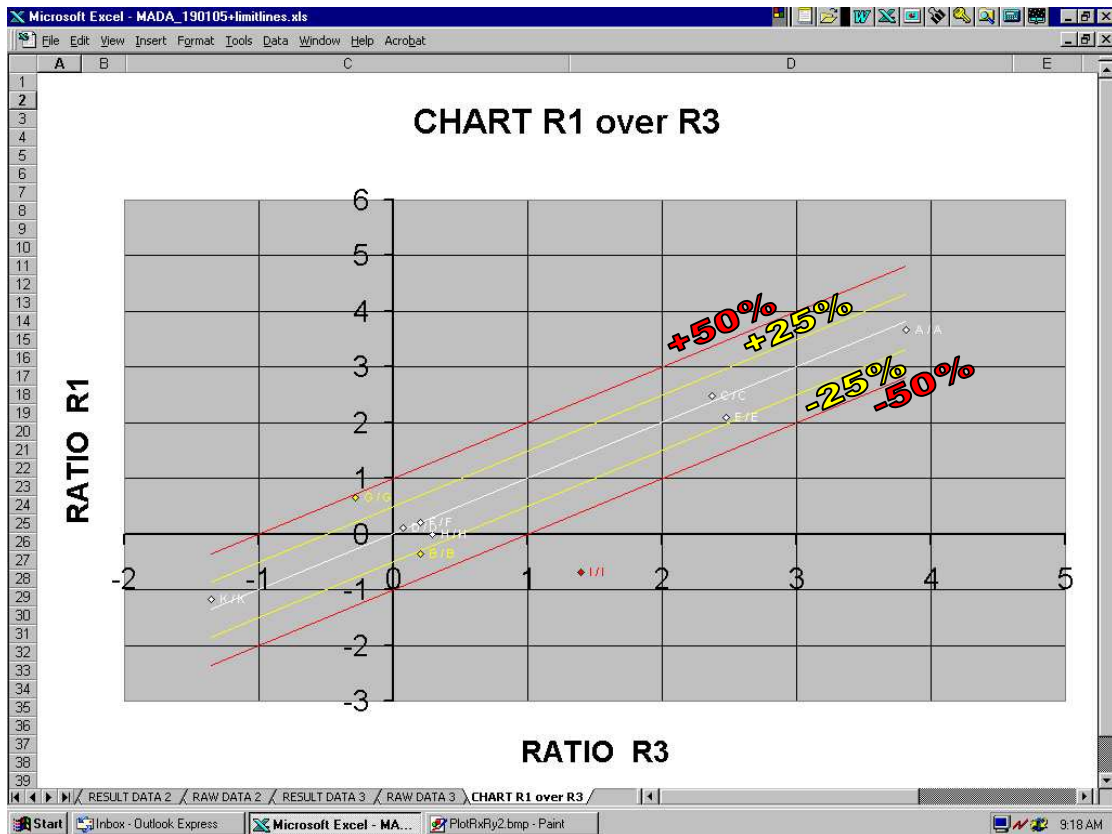


Figure 48: CHART Ry over Rx

5 Save Data

There are two options implemented to save data which are exported from MADA. They are offered after pressing the 'Save Data' button.

a) Save a collection of MADA worksheets to a new Excel workbook.

Worksheets containing 'RAW DATA', 'RESULT DATA', or a 'CHART' can be saved into a new excel workbook.

Additionally the 'System log' can be included and is stored on a separate worksheet.

Press 'Save Data' to open the dialog box. Select the worksheets to save, and edit the filename and location if desired.

The default file name for saving is:

MADA_RESULTS_[date]_[time].xls

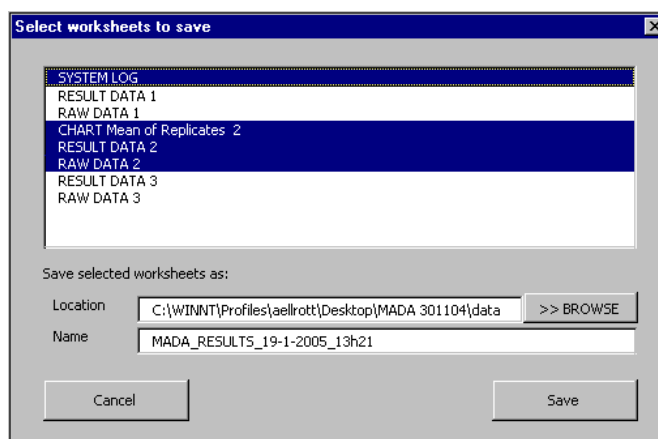


Figure 49: 'Save data' dialog

NOTE: If the selection contains a lot of worksheets, especially chart worksheets, it can take Excel some minutes to copy all the selected worksheets into a new workbook for saving.

b) Extract and save selected data to an export file.

This option allows to create a single export file in text or Excel format by combination of selected information from different 'RAW DATA' and 'RESULT DATA' files. The user can select the data columns, which are saved in the export file. The header can contain free text information or an automatically generated abstract of the system log file.

The 'TRIM DATA SET' option allows to remove 'empty' data lines automatically.

With the 'PREVIEW EXPORT DATA' option the selected data will be first shown in a new Excel worksheet to inspect or modify the data set, e.g., to manually insert a column which combines results of different worksheets by calculating the average value.

The key combination 'CTRL+SHIFT+E' will start the final export after the preview and modification.

The default file name for export is: MADA_EXPORT_[date]_[time].[extension]

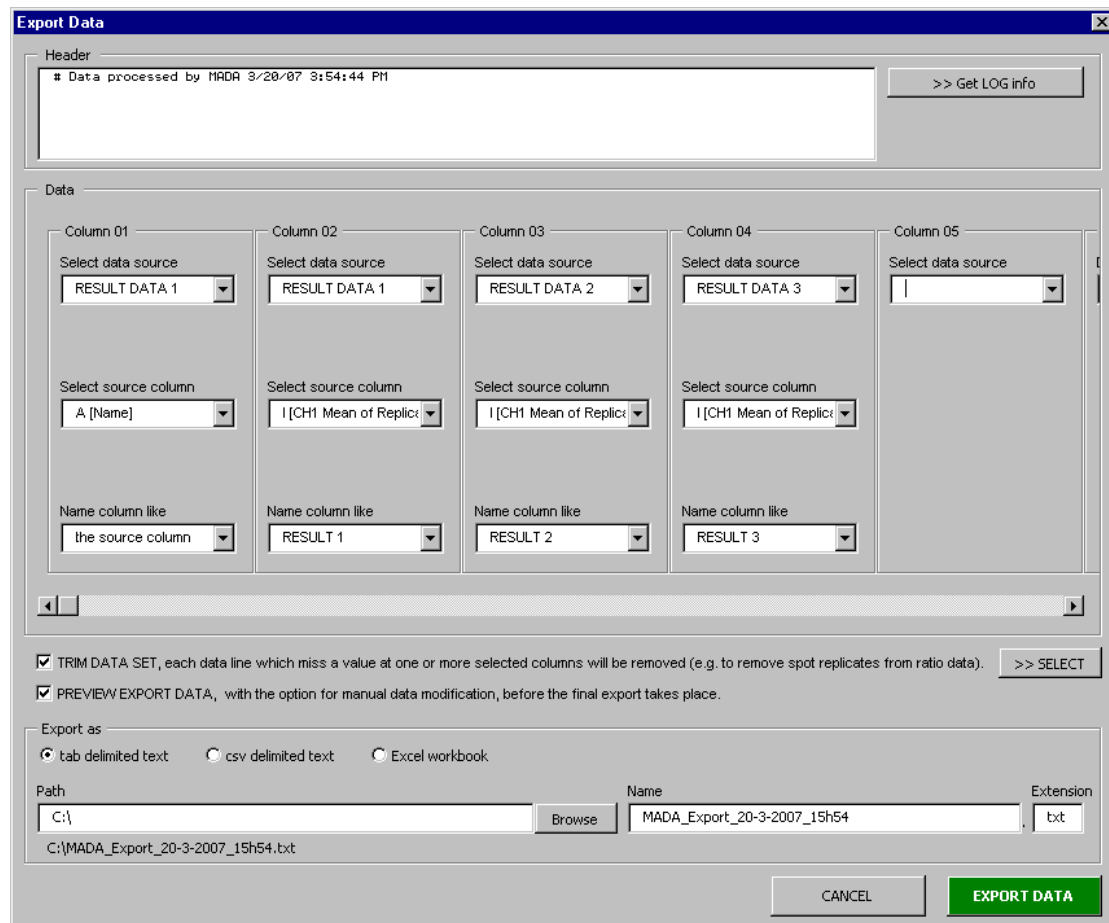


Figure 50: 'Export data' dialog

6 Remove data sheets

Use the 'Remove data sheets' dialog to delete worksheets or charts from MADA. Multiple selections are possible.

It is also possible to delete a worksheet using a standard Excel function. Right click on the worksheet name tab and select 'delete' from the context menu. **Note:** NEVER remove the 'COMMAND' worksheet.

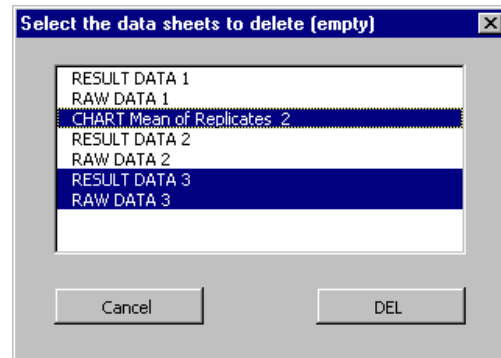


Figure 51: 'Remove data sheets' dialog

This page is intentionally left blank to support double-sided printouts.

7 System log

Central steps of data analysis such as calculations, as well as the corresponding settings and errors, are recorded by MADA. Press 'Show system log' for visualisation. Use 'SAVE' for storing information as a text file. Information will not be stored automatically and are reset with each restart of MADA. Manual reset is possible via the 'CLEAR' button.

Note: The system log can be stored as excel worksheet using the main windows 'Save Data' option.

```

LOG
*****
MADA started at: 4/11/05 10:28:00 AM
*****
IMPORT    RAW DATA 1      4/11/05  10:28:18 AM  C:\data\ScanArray0250.csv
IMPORT    RAW DATA 2      4/11/05  10:28:18 AM  C:\data\ScanArray1000.csv
EXTRACT   RAW to RESULT 1   4/11/05  10:28:41 AM      Ch2=YES  Ch3=NO
EXTRACT   RAW to RESULT 2   4/11/05  10:28:52 AM      Ch2=YES  Ch3=NO
CALCULATE RESULT DATA 1   4/11/05  10:29:36 AM    CH1,CH2
-----
CHx Local Background Correction                                [CH1,CH2]
CHx is Signal? [CHx > Chx Backg. + 2 * Chx Backg. StDev]    [CH1,CH2]
+RESULT-> CH1: 35 of 250 spot signals are insignificant <14%!
+RESULT-> CH2: 71 of 250 spot signals are insignificant <28%!
-----
CHx has Outlier? [Deviation difference > 50%]                [CH1,CH2]
+RESULT-> CH1: 30 of 215 significant spots are outlier <13%!
+RESULT-> CH2: 28 of 179 significant spots are outlier <15%!
-----
CHx Mean of background corrected replicates [min. 2 valid signals] [CH1,CH2]
+RESULT-> CH1: 8 of 72 mean of replicates are invaild <11%!
+RESULT-> CH2: 17 of 72 mean of replicates are invaild <23%!
+> Stdev over 'mean of background corrected signals'
+> Normalization of CHx at mean of overall signal [Basis CH2]
-----
Intensity & Ratio                                           [Basis CH2]
+RESULT-> 19 of 72 data points could not be calculated <26%!
IMPORT    RAW DATA 3      4/11/05  10:32:14 AM  C:\data\Test.csv
EXTRACT   RAW to RESULT 3   Gene names: ERROR: Block Start not found!
-----
SAVE      CLOSE      CLEAR

```

Figure 52: System log window

This page is intentionally left blank to support double-sided printouts.

8 Installation and Setup

MADA is using visual basic for application, coming along as a macro embedded in an Excel workbook.

Microsoft Excel 97 or higher must be installed on a windows-based operating system together with the add-in for VBA support. Normally the later is automatically installed by the standard installation routine of Excel® or can be obtained by using the installer from the Microsoft® Office® CD-Rom. Compare *System requirements and Performance* for further information on system requirements.

8.1 Installing MADA on your PC

Just copy the MADA folder including all files to your hard disk drive and open the MADA.xls file with Microsoft Excel to start, like described at: *Installation 'by hand'*. To prevent from problems caused by Excel, the *Automatic installation* is recommended.

If you run Excel 2000 or higher it is necessary to setup Excels security settings after the installation. See the following section *Setup Excel to run MADA*.

Installation 'by hand':

Download the compressed file **MADA.zip** from <http://www.megx.net/mada> and extract it to a folder on your hard disk drive. In this folder, you will then find MADA.xls together with some extraction filter files (.mef). Close all currently open Excel applications and double-click on MADA.xls to start. **Note:** Fast lowess calculation will only work if the module 'MADALowessCalcModule.exe' is present in this folder and the folder is not write protected.

Depending on the current Excel installation, object libraries or dll's required by MADA can miss. MADA will check for them during load.

An error is produced if they are missing.

Note: This message can also occur if another Excel application is already open. It is strongly recommended to close Excel before opening MADA.

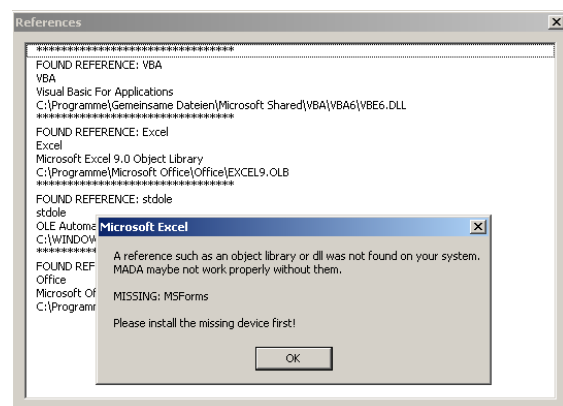


Figure 53: Error caused by missing reference files

Automatic installation:

We provide an installation and setup package for automatic installation. The package includes all components required by MADA such as object libraries, dll's, extraction filters, the speed-up module for fast lowess calculation as well as the MADA manual. Download and execute the file

SetupMADA.exe from <http://www.megx.net/mada> and start the installation procedure by double-click. Please read the information which shows up during the installation process, carefully. Afterwards, close all open Excel applications and double-click on the MADA icon on the desktop or follow the link in the start menu.

8.2 Setup Excel to run MADA

The Visual Basic for Applications add-in to support the macro functionality is required. Although it is normally part of every standard installation of Excel it might be missing or deactivated in some cases. On problems run the installer from the Microsoft® Office® CD-Rom, choose the user defined installation mode and select Visual Basic for Applications (VBA) to install.

Due to the increasing problems caused by computer viruses, Microsoft has implemented some protection against potential macro viruses in the newer Excel versions. For instance, Excel 2000 macros are disabled by default and the access of Excel 2002 (Office XP) to Visual Basic projects is protected. Therefore, the corresponding Excel settings must be changed before starting MADA.

a) Enable macros

If execution of macros is disabled, the following error will occur during MADA start:

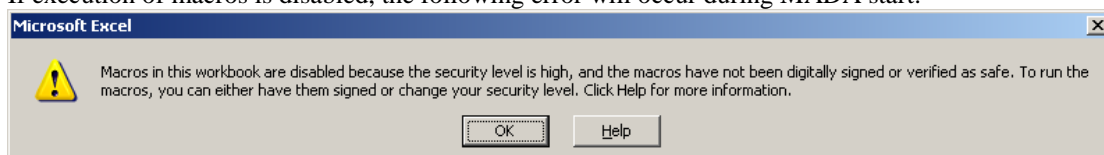


Figure 54: Error, macros disabled

Open Excel and select *Tools > Macro > Security*. Change the security level to 'Medium' or 'Low'.

If the security level 'Low' is selected, Excel will not ask for execution of MADA macros anymore. However, please note that this is also the case for any other Excel macro!

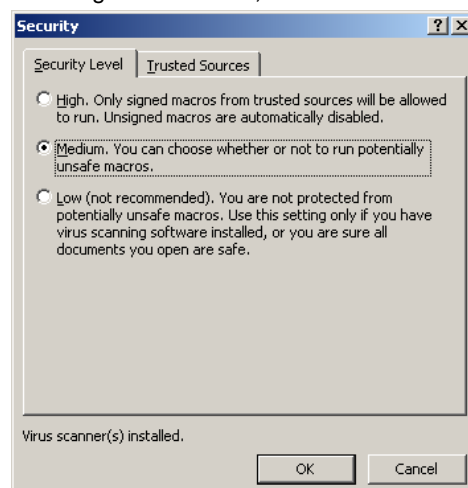


Figure 55: Select security level

If the security level 'Medium' is selected, Excel will ask on every start of MADA (or any other macro) whether macros should be disabled or enabled.

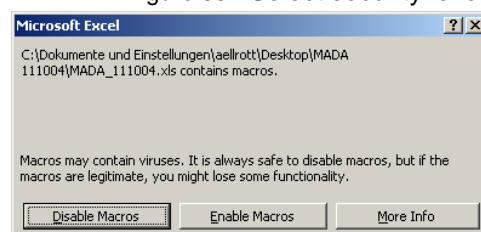


Figure 56: Medium level start question

b) Enable access to Visual Basic projects

If access to Visual Basic projects is prohibited an error message comes up during start of MADA.

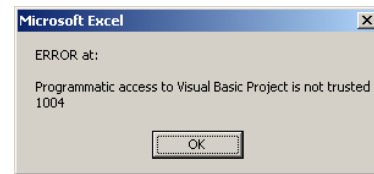


Figure 57: Error, VB project access

In Excel, go to *Tools > Macro > Security* and select Visual Basic projects as trusted source.

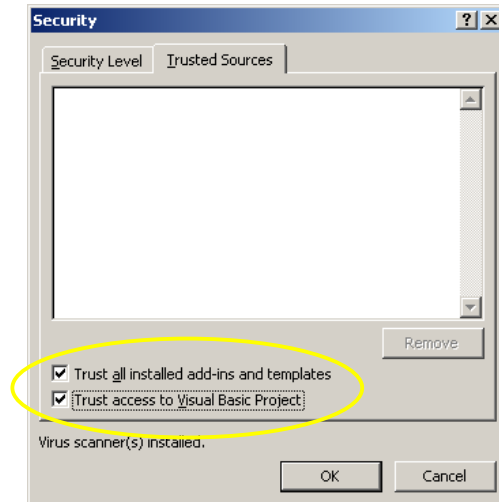


Figure 58: Enable VB project access

This page is intentionally left blank to support double-sided printouts.

9 System requirements and performance

MADA is written as a Visual Basic for Application (VBA) Macro using Microsoft Excel. To run the program, Microsoft Excel '97 or higher and the add-in for VBA support must be installed on a Windows operating system.

Note: Excel on Mac OS is currently not supported.

The calculation speed of MADA is mainly limited by Excel because it has to parse and interpret the VBA program code during runtime and the interaction with Excel, e.g., the readout of Excel cells takes a significant amount of time. For MADA 2.0 the algorithms were revised in terms of calculation speed, which leads to a much better performance compared to the older versions of MADA.

Although MADA is fully functional without any additional software tools, a special module written in Visual Basic 6 is provided, which will fasten the work load intensive lowess calculation if the module is present in the MADA program folder. This module holds the same lowess algorithm like the one directly implemented in MADA, but as the module is a compiled executable file it can run much faster than the macrocode of Excel. This module is already included in the standard installation of MADA.

MADA can only handle data files that contain a maximum number of 60,000 lines, each representing the description of a single spot for a maximum of three different fluorophores/channels. The files must be in the delimited text or the Excel format.

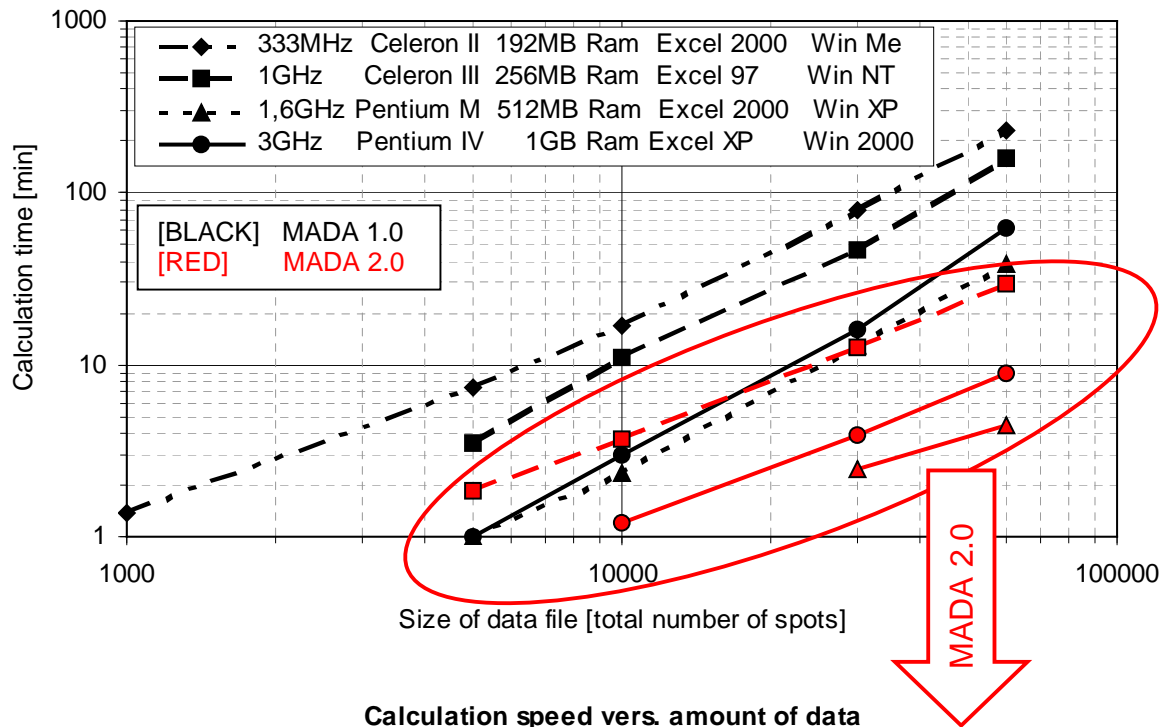
MADA 2.0 can now process up to 24 of this data files, but be aware that each imported file leads to one raw data worksheet, one result data worksheet, plus additional chart worksheets. This can easily blow up the file size of the Excel workbook to several 100MB and therefore slow down Excel.

Note: Excel 97 has known problems in handling workbooks containing more than 60 worksheets.

MADA was optimised for a screen resolution of 1024 x 768 and can not be used with lower resolutions.

For the performance tests, we created artificial 2-channel data files holding different numbers of data points from 250 up to 60,000. The data are repeated every 228 lines, so that the percental calculation results are nearly independent from the final file size. Every of the 228 lines data block consist of 62 'genes' in triplicates, one 'gene' in 6x replication and four 'genes' in 9x replication. MADA's spot signal significance test assigns 14% of the spots as insignificant at channel 1 and 28% insignificant at channel 2 using the $CHx > Chx \text{ Backg.} + 2 * Chx \text{ Backg. StDev}$ criteria. The outlier test (50%) finds additional 13% outlier at channel 1 and 15% outlier at channel 2.

The performance of MADA 2.0 was tested with different hardware configurations and is significantly increased compared to the performance of MADA 1.0 like shown in the following chart.



Calculation speed vers. amount of data using MADA 2.0 standard settings with background correction, signal significance test, outlier test and mean of overall signal normalization done on different computer hardware with different software combinations

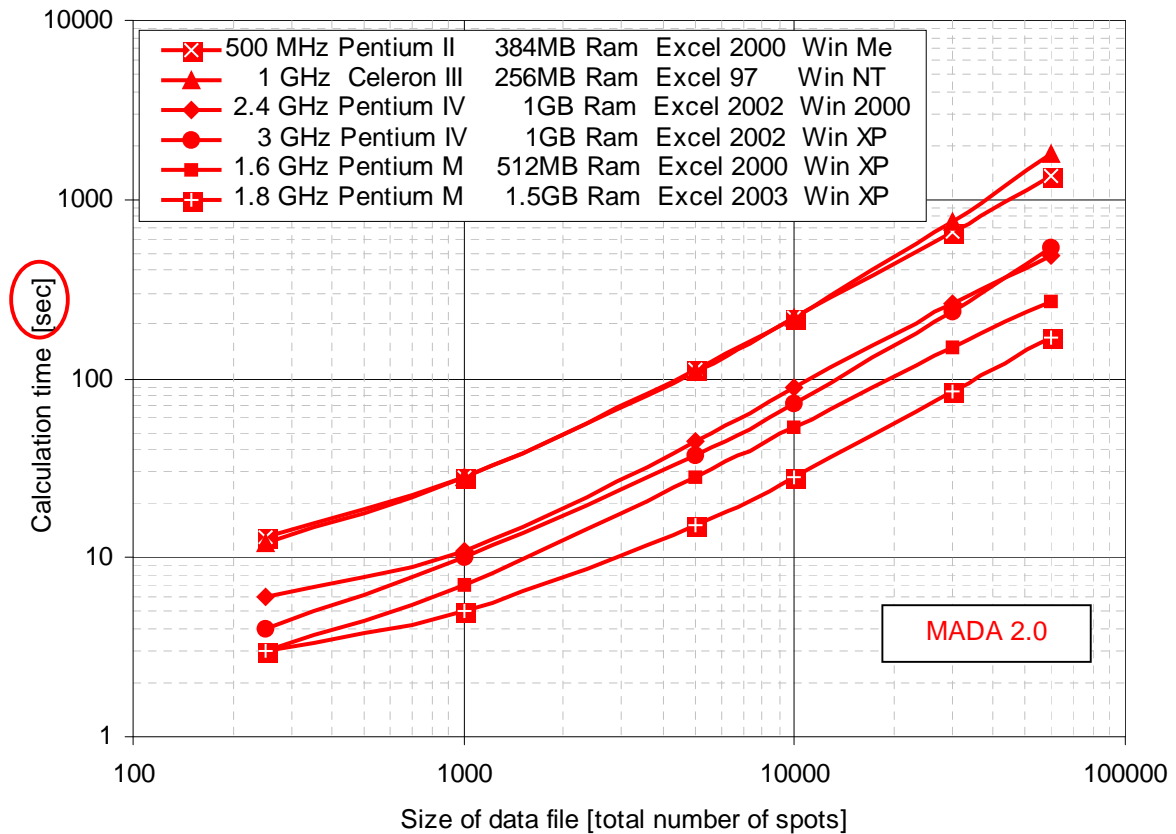


Figure 59: Performance of the MADA calculation step

With limited data sets representing 5,000 spots, MADA is performing well even if doing calculations on older machines. For larger data sets, especially if lowess normalisation is applied, faster machines with at least 2GHz and 512 MB of RAM are recommended.

The most workload is produced by the MADA outlier test and lowess normalisation. For lowess, an additional module is provided which significantly speeds up the calculation. The following chart shows how the performance is influenced by this module and in combination with the outlier test.

**Calculation speed vers. amount of data
using different combinations of more or less workload intensive methods
done on a 1.6 GHz Pentium M with 512 MB Ram using Excel 2000 at Win XP**

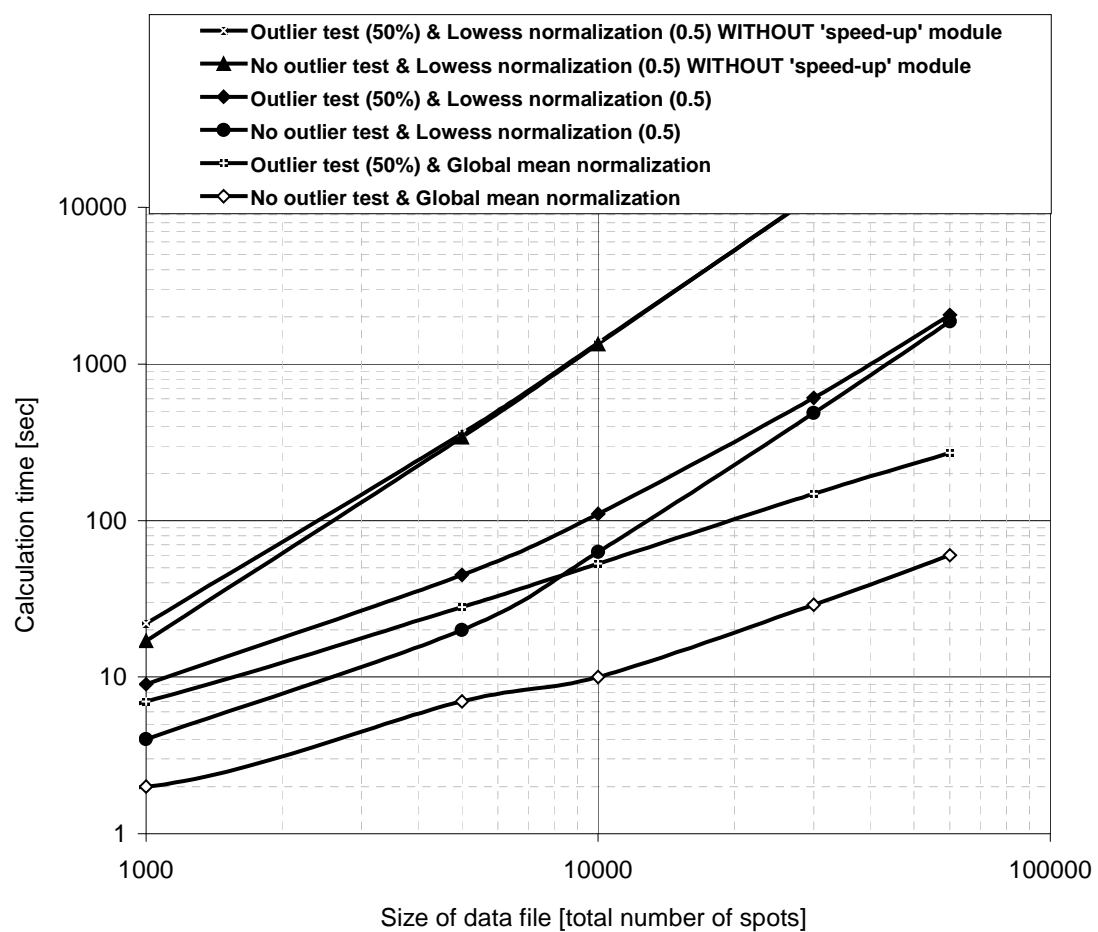
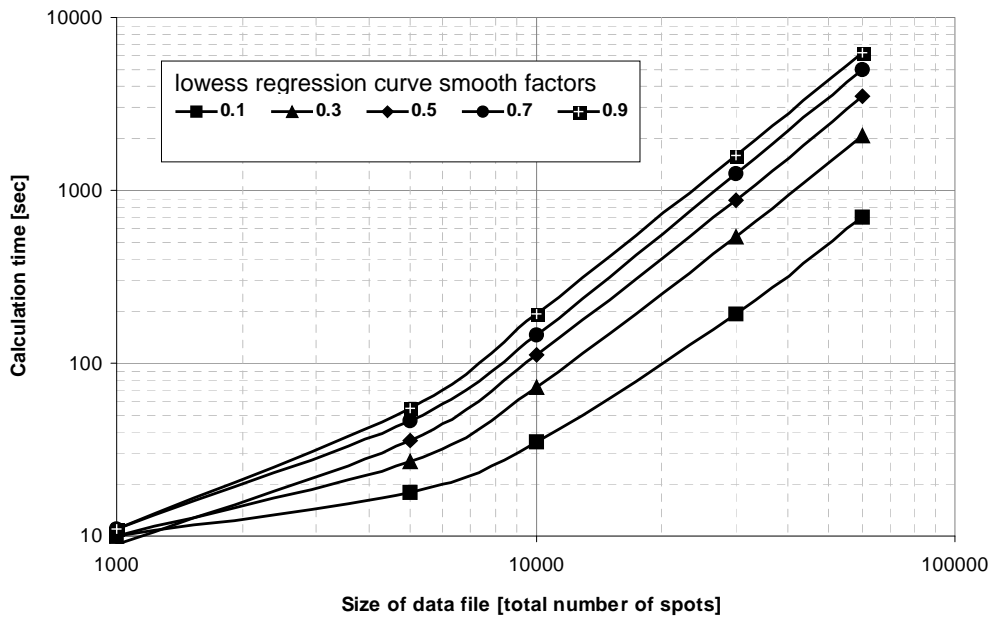


Figure 60: Performance of different work load intensive methods

The performance of the lowess normalisation also depends on the smooth factor selected. The higher the smooth factor, the longer it takes to calculate the lowess regression values. For microarray applications usually a smooth factor within the range of 0.25 to 0.5 is applied. The following charts are based on calculations using the MADA lowess 'speed-up' module. They show that data files up to 10,000 data points can be processed even on older machines within a reasonable time frame but they also show that for large data sets with more than 30,000 data points even faster machine requires much

**Calculation speed vers. amount of data
using lowess normalization with different smooth factors
done on a 1 GHz Celeron III with 256 MB Ram using Excel 97 under Win NT**



**Calculation speed vers. amount of data
using lowess normalization with different smooth factors
done on a 1.6 GHz Pentium M with 512 MB Ram using Excel 2000 at Win XP**

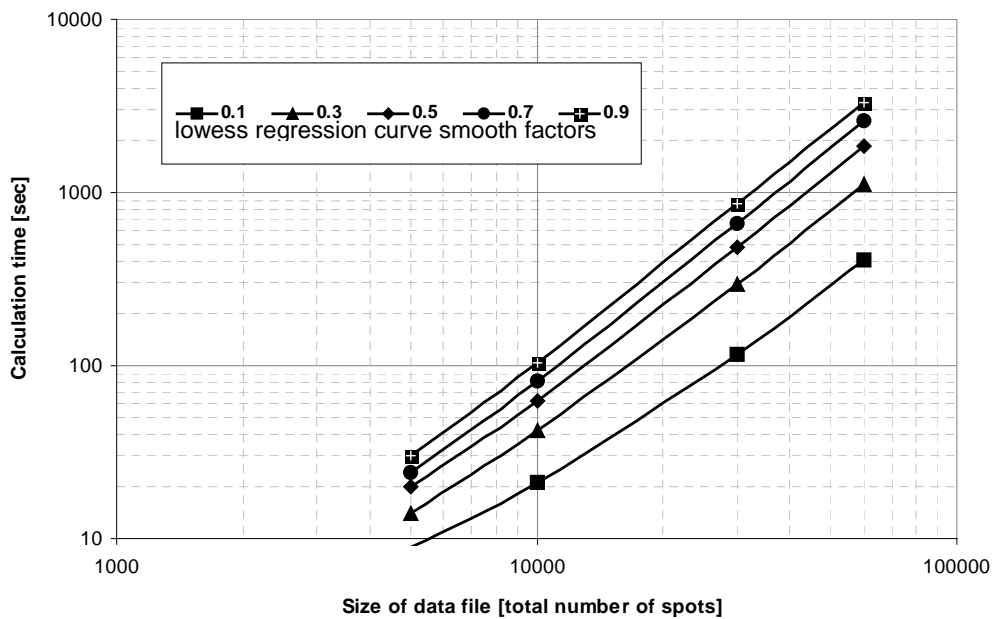


Figure 61: Lowess performance with different smooth factors

10 Troubleshooting

We have done our best for programming MADA without any bugs or incompatibilities but we can not guarantee for the complete absence of any problems. Here, you can find some explanations and solutions for the most common errors. For further information and our contact address, please refer to our website <http://www.megx.net/mada>.

a) General

A message referring to macros pops up at MADA start.

Excels security setting must allow macros because MADA is implemented as VBA macro.

→ Select 'Enable Macros' if ask and read chapter: *Setup Excel to run MADA*

Error 1004: Programmatic access to Visual Basic Project is not trusted, pops up at MADA start.

→ Enable access to Visual Basic projects like described at: *Setup Excel to run MADA*

A message pops up that a reference to an object library or dll is missing.

An object library or dll required by MADA is not available on your PC or currently blocked by another application.

→ Make sure that all other Excel applications and worksheets are closed before starting MADA.

→ Update your system with the references required, see: *Installing MADA on your PC*

b) Import Raw Data

An error message pops up during import.

Avoid the use of special characters or white spaces in the data file name.

The whole data path should not exceed more than 128 characters.

The import of a delimited file sticks and no or corrupted data are imported.

Some programs create data files including 'strange' characters so that the files can not be opened with Excel or leading to corrupted data.

→ Delete these characters before data import by using a text editor such as Notepad.

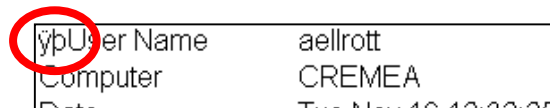


Figure 62: Example for a 'strange' character

After import, some names have been misspelled/changed.

It is a known problem, that Excel could misinterpret identifiers, e.g. the gene name DEC-1 is interpreted as date and automatically changed to 1-Dec during import. [Zeeberg et al. 2004]

After importing a delimited file, the values differ entirely from the original data or all data were extracted into a single column.

The list delimiter and/or decimal separator was interpreted not correctly because the delimiters used in the data file does not correspond to the ones of the operating system.

The standard list delimiter, decimal separator, and digit group symbol are based on the regional settings (country code) of your operating system.

	List delimiter	Decimal separator	Digit group symbol
US Windows	, (comma)	. (<i>dot</i>)	, (comma)
German Windows	; (semicolon)	, (comma)	. (<i>dot</i>)

Figure 63: Examples for default regional settings of different OS

If you run Excel on a German OS, the import of a delimited file created on a US OS will very likely lead to wrong interpretation.

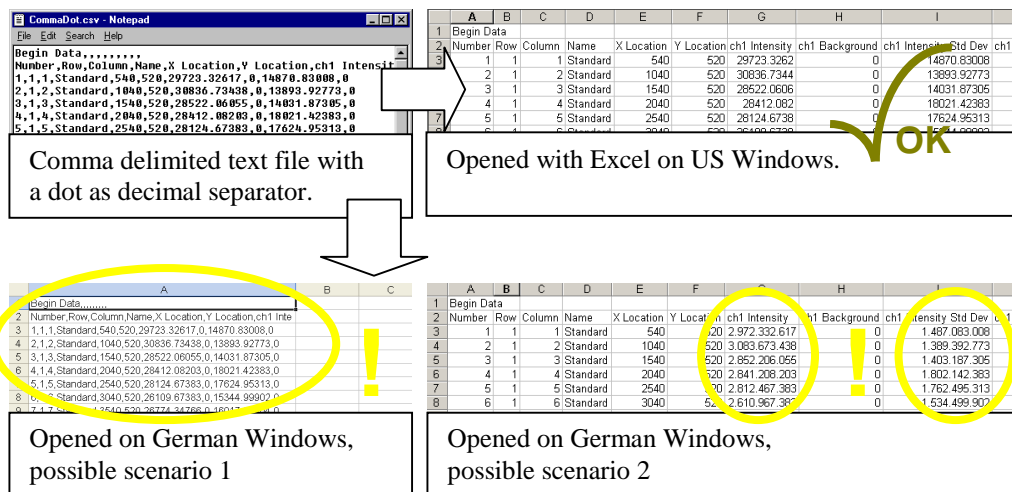


Figure 64: Possible problems with delimited files on different OS

→ Try the 'Check and convert non xls files ...' option to check for system compatibility automatically with the option to create a compatible file copy if needed or change the regional settings of your operating system according to the delimiters used in the file to import.

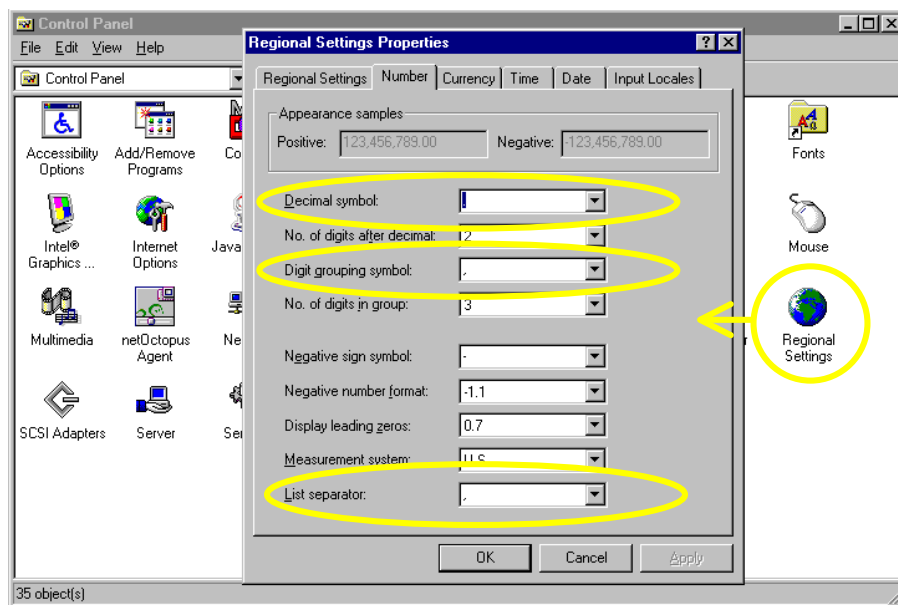


Figure 65: Regional settings on a US version of WinNT

The settings can be found at: 'control panel' → 'regional settings', of your operating system.

Note: With Excel 2002 (Office XP) it is now possible to use a different decimal separator as the OS defines it. Select: 'Options' → 'International' → 'Number handling' in Excel.

Unfortunately, problems with different list separators can not be solved here.

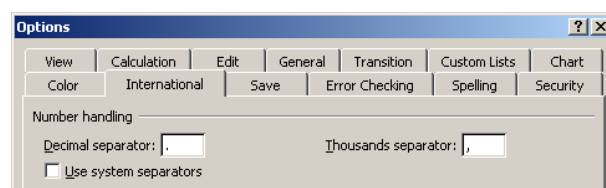


Figure 66: Excel 2002, number handling

c) Extract Data

Error message during data extraction.

Identifiers, prompted variables, or data columns defined in the extraction filter were not found at the raw data file.

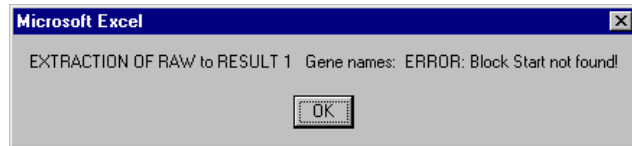


Figure 67: Example for error during extraction

- Assure that the raw data were correctly imported. Otherwise refer to: *Troubleshooting* → *Import Raw Data*
- Assure that the appropriate extraction filter was selected and check if the filter settings match to the raw data organisation. Read the chapter *Extraction filter* for more detailed information.
- Check the raw data if all parameters and channels, which should be extracted, are really exists.
- Are variables prompted correctly typed?

d) Calculation

Error message 'formulalength' pops up during calculation.

A formula in Excel is limited to a length of 1024 characters, including hidden parameters such as the actual path of the workbook.

After start of calculation, MADA tests the maximal possible formula length for your system. The error occurs if a formula reaches

this limit. This could happen at high numbers of replicates. Sometimes the reason for high replicate numbers are empty spots or control spots all named the same. With large datasets, this limit is reached much faster than on small datasets due to the longer cell reference numbers:

- ='{path}{MADA.xls}{sheet}'AVERAGE(E1,E2,E3,E4)
- ='{path}{MADA.xls}{sheet}'AVERAGE(AA12001,AA12002,AA12003,AA12004)

→ Different probes should be named different. Delete empty spots from the data set or give them different names. Try to install MADA in the root directory, to shorten the actual file path.

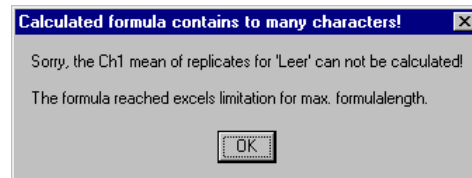


Figure 68: Calculation error

Outliers are expected but were not found.

There must be a minimum of three replicates for each probe within a single dataset to perform an outlier test, and at least one more than the half of all replicates must be assigned as significant.

→ Reduce the deviation difference of the outlier test to increase sensitivity of the test.

The mean of replicate value is zero.

There are more outliers or signals not significantly above the background as defined by the 'Minimum number of replicates that must be assigned as signal'.

→ Adjust the value for 'Minimum number of replicates that must be assigned as signal'.

Note: It is recommended to choose a value according to this formula:

$(\text{number of replicates} / 2) + 1$

- Adjust the parameter for the significance test. (default = 2)
- Increase the deviation difference at the outlier test.

**The normalised mean of replicate value is zero or division by zero.
The ratio or intensity values are zero or division by zero.**

The corresponding mean of replicate is zero. Refer to: *The mean of replicate value is zero.*

Calculated results are incorrect.

Raw data values have been imported in the wrong number format, so Excel interprets their decimal symbol in a wrong way or they are seen as text string.

- The delimiters used in the raw data file must correspond to the ones of the operating system. Refer to: *After importing a delimited file, the values differ entirely from the original data or all data were extracted into a single column.*
- Numbers in a text file should not have quotation marks. This could lead to an interpretation as text string, Excel could not calculate with. Delete the quotation marks in the raw data file using a tool like Notepad, then import, extract and calculate again.

e) Plot Chart

There are expected data points missing in the chart.

- Plot again and make sure that the missing data points are available and activated in the probe selection box, which appears after the worksheet selection.
- A value for the data point was not calculated because too many signals were defined as not significant or an outlier. Check *Troubleshooting* → *Calculation* for further information

Names are repeated in the list box and charts.

Labelling and / or colouring do not work properly.

- There are data without a name in the corresponding row of the dataset. MADA then interprets the row as end of data set, causing internal problems. Name all data rows in the raw data file, then extract, calculate and plot again.

Bisecting line and / or axis scale is corrupted. (Zigzag lines, very large scale)

- The internal number format of VBA does not fit to the format expected by Excel. Try to change the operating systems regional settings to 'dot' as decimal separator and 'comma' for the digit group. On Excel 2002 use the similar settings at the number handling option. Check *Troubleshooting* → *Import Raw Data* for further information on regional settings.

Data points in a 'ratio over ratio plot' seem to fall on a virtually 90° rotated bisecting line.

- Potentially, the assignment of channels to corresponding fluorescent dye is exchanged in the raw data set or the assignment of the channel for ratio calculation is wrong. Choose the correct ratio calculation channel assignment according to the raw data set, calculate and plot again.

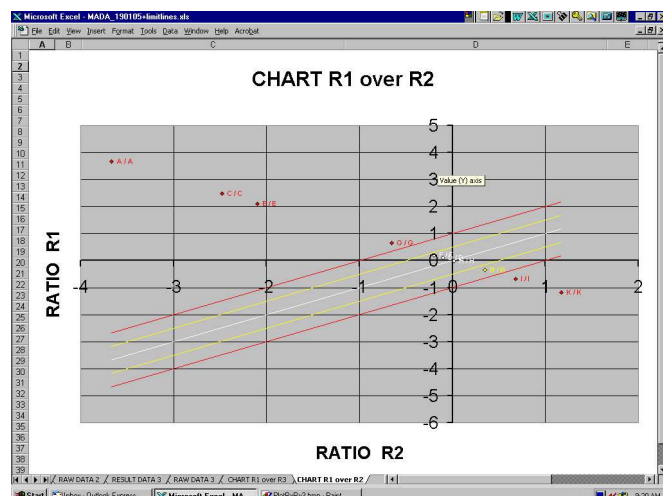


Figure 69: Ry-Rx-plot, assignment of wrong channels

This page is intentionally left blank to support double-sided printouts.

Appendix

Acknowledgments

Thanks to:

- Chris Würdemann who came up with the initial idea of MADA, contributed to the concept and did most of the testing.
- Frank Oliver Glöckner for supporting the project.

References, resources and further reading

- [ABOUT Visual Basic] Harald M. Genauck, ABOUT Visual Basic, <http://www.aboutvb.de>
- [ActiveVB] ActiveVB-Team, ActiveVB, <http://www.activevb.de>
- [Bortz] J. Bortz, Lehrbuch der Statistik für Sozialwissenschaftler. Springer-Verlag Berlin (1977, 1979)
- [Forster et al. 2003] T. Forster, D. Roy and P. Ghazal, Experiments using microarray technology: limitations and standard operating procedures. *Journal of Endocrinology* 178:195-204 (2003)
- [Kroll et al. 2002] T. C. Kroll and S. Wölfl, Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. *Nucleic Acids Res.* 30: e50 (2002)
- [MathWorld] Eric W. Weisstein, MathWorld - A Wolfram Web Resource, <http://mathworld.wolfram.com>
- [NIST/SEMATECH] NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook>
- [Quackenbush 2002] J. Quackenbush, Microarray data normalization and transformation, *Nature Genetics* 32, 496-501 (2002)
- [VB-fun] Detlev Schubert, VB-fun.de, <http://www.vb-fun.de>
- [Xlimits] Xlimits, <http://www.xlam.ch/xlimits/index.htm>
- [Zeeberg et al. 2004] Zeeberg B.R., Riss J., Kane D.W., Bussey K. J., Uchio E., Linehan W. M., Barrett J. C. and Weinstein J. N. Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics* 5:80 (2004)

This page is intentionally left blank to support double-sided printouts.

Freeware License Agreement

This software is provided "as-is," without any express or implied warranty. In no event shall the author be held liable for any damages arising from the use of this software.

Permission is granted to anyone to use this software for any purpose and to redistribute it, provided that the following conditions are met:

1. All redistributions must retain all copyright notices that are currently in place, and this list of conditions without modification.
2. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software to distribute a product, an acknowledgment in the product documentation would be appreciated but is not required.
3. Modified versions must be plainly marked as such, and must not be misrepresented as being the original software.

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

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

(c) 05.2007