



# ENM TUTORIALS

How to use the AffyQC web tool of ArrayAnalysis.org for quality control and pre-processing of Affymetrix microarray data

RELEASE DATE:	May 23rd 2016
USE:	How to use the AffyQC web tool of ArrayAnalysis.org for quality control and pre-processing of Affymetrix microarray data
VERSION:	V.1.0.
MAIN AUTHOR:	Friederike Ehrhart
PARTNER:	UM
CONTACT DETAILS:	<a href="mailto:friederike.ehrhart@maastrichtuniversity.nl">friederike.ehrhart@maastrichtuniversity.nl</a> <a href="mailto:linda.rieswijk@maastrichtuniversity.nl">linda.rieswijk@maastrichtuniversity.nl</a> <a href="mailto:egon.willighagen@maastrichtuniversity.nl">egon.willighagen@maastrichtuniversity.nl</a> +31(0)43-38 82913
AUTHORS:	Lars Eijssen, Anwasha Bohler, Linda Rieswijk, Egon Willighagen, Penny Nymark
LICENCE:	CC-BY 4.0





# TABLE OF CONTENTS

[1. INTRODUCTION](#)

[2. APPLICATION DETAILS](#)

[First step: load the CEL files](#)

[Second step: describe the dataset](#)

[Third step: define your analysis](#)

[Execution step](#)

[Getting the results](#)

[3. ACKNOWLEDGMENTS](#)

[4. REFERENCES](#)

[5. KEYWORDS](#)

# 1. INTRODUCTION

[ArrayAnalysis.org](http://ArrayAnalysis.org) is an open source, free to use online platform for analysis of microarray data - and an alternative program for [Chipster](#) (tutorial also available). ArrayAnalysis is a webtool, so there is no need for download or access code, and it provides more extensive quality control than Chipster but it is limited to two microarray formats: Affymetrix and Illumina. The exact microarray type (e.g. Affy-1) is automatically recognised. This tutorial shows how to use the Affymetrix quality control (affyQC) module which is designed for doing quality control and preprocessing of microarray data from Affymetrix microchips. All source code has been written in R and is open-source, available under the [Apache License version 2.0](#). It is available on our [Download](#) page.

affyQC can be run :

- on-line via the [arrayanalysis.org](http://arrayanalysis.org) webportal (follow "[Get started](#)").
- locally as an automated R workflow provided via a wrapper function

The main functions of affyQC are:

- to compute array quality information;
- to plot images that allow identifying any aberrations present in the dataset;
- to return pre-processed data and QC reports.

**Bug tracking system:** If you encounter an issue by using the code, you can report it at any moment [by email](#) or, once you have your own account, using our [internal tracking system](#). You can also use this system to post comments or suggest features.

**Example datasets:** Note that three example datasets has been made available on our [Download](#) page. They include:

- dataset raw .CEL files,
- description file,
- affyQC output files:
  - execution logfile,
  - report file (PDF),
  - zip archive with images and tables and
  - normalised data (text file)

You can access the on-line module on [arrayanalysis.org](http://arrayanalysis.org) webportal: follow "[Get started](#)".

**JavaScript** has to be enabled (activated) in your web browser. You will be warned if it is not the case. You can activate it at any time in the browser options (see [activatejavascript.org](http://activatejavascript.org) if needed)

You don't need to log in; you just need to prepare a zipped file containing all your Affymetrix .CEL files and possibly a file describing your dataset, called the description file. A presentation of this description file is available in the fourth section, subsection "[Parameter description](#)". The on-line module contains three steps before the launch of the analysis:

- [Step1](#): First you load the archive of .CEL files
- [Step2](#): Then you complete the description of the dataset
- [Step3](#): And finally you choose the plots to be computed and their parameters.

Then:

- [Execution](#): The module is executed with the settings you choose
- [Results](#): You get the results after the execution step, or by e-mail.

## 2. APPLICATION DETAILS

### FIRST STEP: LOAD THE CEL FILES

The following picture shows the screen for the first step:

[Get started](#) [Download sources](#) [QC Modules description](#) [Documentation](#) [Bug tracker](#)

## Run ArrayAnalysis!

### QC & pre-processing of Affymetrix expression chips

*Before running this module, you may visit its referred [user guide](#)*

Enter your Affymetrix raw dataset: a zipped file containing the .CEL files of your arrays

No file selected.

[Or discover \[Affymetrix QC & pre-processing\] using an example dataset \(Example1\)](#)

*Please don't make changes or click any button while data is uploading*

The interrogation mark button will help you by giving you a contextual help. Note that this feature is available when Javascript is activated and is not yet supported by GoogleChrome and Safari browsers. Loading the zip file may take a while as CEL files are heavy; don't click any button after clicking on the "Next" button otherwise the loading of the file may be compromised. When the file is loaded without error, you are automatically directed to the next step. Otherwise you get a message indicating the error encountered:

### There were some errors in the input form:

The zip file is corrupted (file transfer was interrupted)

[Go back to the input form](#)

## SECOND STEP: DESCRIBE THE DATASET

The following picture shows the screen obtained after completing the first step:

Get started
Download sources
QC Modules description
Documentation
Bug tracker

**[QC & pre-processing] Describe your dataset**

Describe your dataset for analysing and coloring the arrays per experimental group: complete the table below or load a description file.

ArrayDataFile	SourceName	FactorValue
GSM789975_MG2009111	Array1	Group1
GSM789974_MG2009111	Array2	Group1
GSM789973_MG2009111	Array3	Group1
GSM789972_MG2009111	Array4	Group1
GSM789971_MG2009111	Array5	Group1
GSM789970_MG2009111	Array6	Group1
GSM789969_MG2009111	Array7	Group1
GSM789968_MG2009111	Array8	Group1

No file selected.
[?]

Reorder samples by experimental group

[?]

*Please don't make changes or click any button while data is uploading*

[Back to the previous step](#)

The interrogation mark buttons will help you by giving you a contextual help. Your dataset has been read and the following information is presented in a three columns table:

Column "ArrayDataFile" contains the .CEL file names of your N arrays found in the input zip file. You cannot edit this column.

Column "SourceName" is filled with Array1 .. ArrayN. These names will be used for the analyses. Feel free to modify these names at the condition you use only unique names.

Column "FactorValue" is always set to "Group1". If you want your array groups to be represented in the analyses and plots, rename the factor groups.

You may also prefer to enter directly this information from a file you have prepared. If this is the case, browse your description file in the second section. If you enter such a file the

information contained in the previous table will be skipped. You'll find a presentation of the description file on the fourth section of this documentation: "[Parameter description](#)"

The last section of the second step form proposes you to reorder the arrays per groups, which is done by default. Thus all the arrays representing the same factor will be grouped together on the plots. If you untick the checkbox, arrays will be ordered as they were in the zip file.

Clicking on the "Next" button will direct to the last step if no error has been detected.

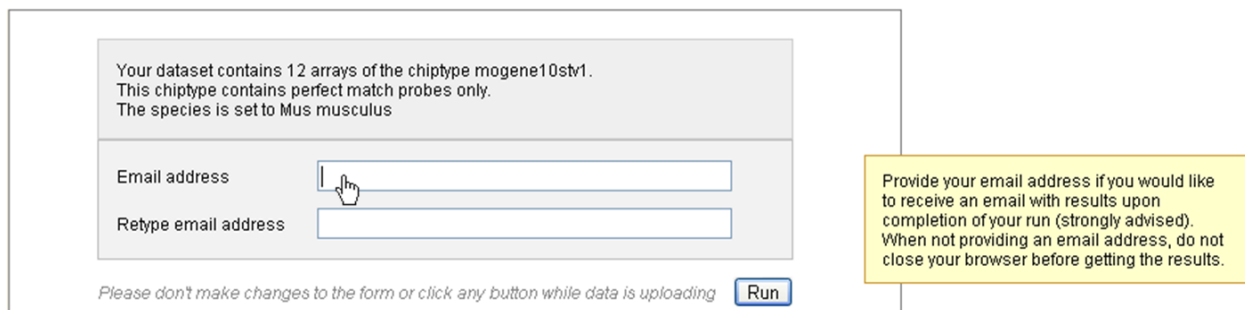
### THIRD STEP: DEFINE YOUR ANALYSIS

The contextual help is not any more given by the interrogation mark buttons: help messages will pop up as soon as you activate a field (for example if you click in a text field or tick a checkbox).

This last input form is divided into three main sections: the first part allows a quick launch, the second part defines in details the analysis parameters applied to the raw data and the last part is dedicated to the pre-processing (parameters for the normalization and re-annotation) and its evaluation (definition of the analysis parameters applied to the normalized data).

#### First part of the input form

The following picture presents the first part; it recalls briefly what your dataset contains and asks you to enter an e-mail. This is optional: if you don't enter your e-mail, you will need to keep the browser opened and not close the page before the end of the calculation. On the contrary, if you enter your e-mail address - which is recommended - you can close the windows as soon as the next page appears and you will be informed of the end of the analysis by e-mail. You would just have to follow the links to the result files given in the e-mail.





You may launch the analysis with the "Run" button right after this first section. In this case default parameters will be used.

Note that if the species was not deduced from the previous step, you will need to fill this field first, or to untick the "Custom annotation" checkbox.

### Second part of the input form

This part contains four frames representing the four families of analysis applied to your raw data: 1) Sample quality, 2) Hybridization and overall signal quality 3) Signal comparability and bias diagnostic and 4) Array correlation.

Most of the parameters are checkboxes that you would tick or untick to indicate whether a certain plot or table has to be computed or not. The analyses and plots are described in the [module description](#) page, which is reachable also from the left vertical menu (we recommend you to open the pages in a new tab to not lose the information entered in the input form you are filling).

Some analyses or plot construction, such as the MA-plot and the hierarchical clustering, need particular parameters. You may modify the default values.

The following picture presents you this part of the input form, which defines the graphs built from the raw data:

Select the plots for the quality control and define their options: [ toggle select all  ]

### Sample quality

Sample prep controls  3'5' ratio  RNA degradation

### Hybridization and overall signal quality

Spike-in controls  Percent present   
 Positive/negative controls  Background intensity

### Signal comparability and bias diagnostic

#### Signal distribution

Scale Factors  Control profiles and affx boxplots   
 Boxplot  Density histogram

#### Intensity-dependent bias

MA-plot  *MA-plot per experimental group or using all arrays*

#### Spatial bias

Array reference layout  Pos/Neg Center of Intensity   
 2D images  All PLM-based images

#### Probe-set homogeneity

NUSE plot  RLE plot

### Array correlation

Correlation plot  PCA analysis  Hierarchical clustering

*Distance calculation method*

*Clustering method*

You may note that all the plots are not selected by default; you may select all of them with the first checkbox: [toggle select all].

You may also note that some plots cannot be selected, such as the "Sample prep controls", the "Background intensity" or the "Scale factors". This is because the dataset used for this example (public dataset available on ArrayExpress: [E-GEOD-13278](#)), was built with PM-only arrays and the construction of these particular graphs uses the [MAS5 algorithm](#) which cannot be applied to PM-only arrays.

Be aware that the generation of 2D PLM-based images for spatial biases are highly time-consuming; the generation of the complete set of images (4 different images representing the raw data, the PLM weights, residuals and residual signs) is not computed by default. See examples of these images on the [description page](#) or on [Bolstad PLM page](#).

### Third part of the input form

The following picture presents the part of the input form concerning the pre-processing step and its evaluation:

Define the pre-processing step and its evaluation:

**Pre-processing: normalization method and annotation**

Normalization method  Normalisation per experimental group or using all arrays

Custom annotation file (CDF)  Annotation type

Species

**Pre-processing: evaluation**

**Signal comparability and bias of normalized intensities**

Boxplot       Density histogram       MA-plot

**Normalized array correlation**

Correlation plot       PCA analysis       Hierarchical clustering

Please don't make changes to the form or click any button while data is uploading.

Use the "Normalization method" drop-down menu to define the pre-processing step. You may chose "none" and keep the raw data. In this case, further parameters will be skipped. By default, the GC-RMA is applied to arrays containing both PM and MM probes and RMA is applied to PM-only arrays.

If the species could have been deduced from the CEL files in the previous steps, the "Species" field is already filled, as shown in this example. Otherwise, you would need to fill this field yourself or to untick the "Custom annotation" checkbox.

Indeed, the probesets will be re-annotated by default, using one of the gene annotation databases (see "Annotation type" drop-down menu) and the "Species" is required for the re-annotation.

After defining the pre-processing, you chose the analyses you want to apply to the normalized data. Only six graphs are proposed (other graphs are not meaningful on normalized data) and the parameters entered for the MA-plot and hierachical clustering applied on raw data will be also used for the normalized data.

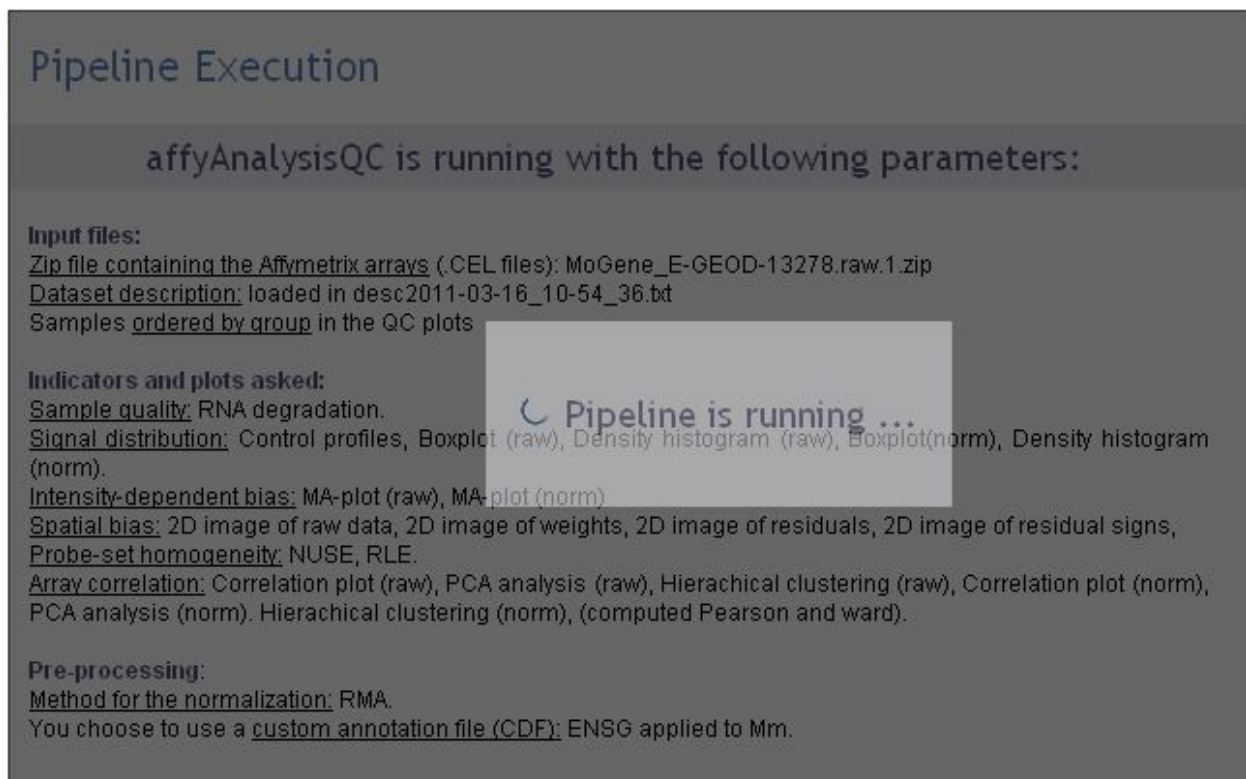
Once the input form is completely filled, you can launch the analysis with the "Run" button. Don't click any button after clicking on the "Run" button and before being automatically redirected to the execution page, otherwise you may compromise your analysis.

## EXECUTION STEP

After the third step, affyQC has all it needs to launch the analysis. The page become grey with a message telling you that the analysis is running. If you entered your e-mail address in the previous step, you can now close the window.

You will find on this page a recalling of the choices you made for this analysis: which files were loaded or created, which plots you decided to create for raw and normalized data and how you managed the pre-processing step.

The following picture shows the screen for the execution step:



**Pipeline Execution**

affyAnalysisQC is running with the following parameters:

**Input files:**  
Zip file containing the Affymetrix arrays (CEL files): MoGene\_E-GEOD-13278.raw.1.zip  
Dataset description: loaded in desc2011-03-16\_10-54\_36.txt  
Samples ordered by group in the QC plots

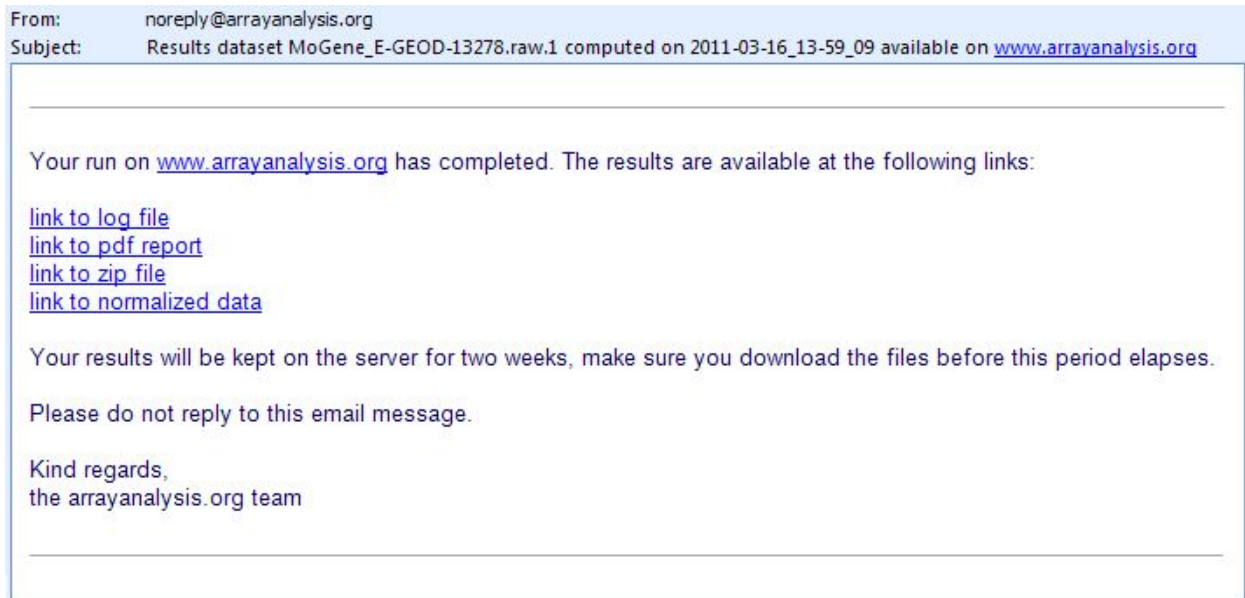
**Indicators and plots asked:**  
Sample quality: RNA degradation.  
Signal distribution: Control profiles, Boxplot (raw), Density histogram (raw), Boxplot(norm), Density histogram (norm).  
Intensity-dependent bias: MA-plot (raw), MA-plot (norm)  
Spatial bias: 2D image of raw data, 2D image of weights, 2D image of residuals, 2D image of residual signs,  
Probe-set homogeneity: NUSE, RLE.  
Array correlation: Correlation plot (raw), PCA analysis (raw), Hierarchical clustering (raw), Correlation plot (norm), PCA analysis (norm). Hierarchical clustering (norm), (computed Pearson and ward).

**Pre-processing:**  
Method for the normalization: RMA.  
 You choose to use a custom annotation file (CDF): ENSG applied to Mm.

Pipeline is running ...

## GETTING THE RESULTS

If you entered your e-mail address during the third step, you will receive an e-mail such as the one presented on the following picture:



The e-mail contains direct links to the log file, PDF report, ZIP file containing the resulting files (png images, usable for your presentations, and result files such as the PMA table) and normalized dataset (presented as a tab-delimited file). If you closed the browser once you analysis was launched, you can only reach these result files through the links given in the e-mail. You cannot access your results from the arrayanalysis.org portal anymore.

On the contrary, if you did not close the browser, the result page presented in the following pictures shows up when the calculation are ended.

A first section gives you the same links to the result files than the e-mail: we recommend you either to save these links or to save the result files because if you did not enter your e-mail, once you close this result page, you will not be able to reach them again.

You can download your result files during one week from the links given by e-mail or by the result page. Make sure you download the files before this period elapses.

This section ends with a frame in which the PDF report is opened. You can visualize the document and save it from this frame.



## Results for MoGene\_E-GEOD-13278.raw.1:

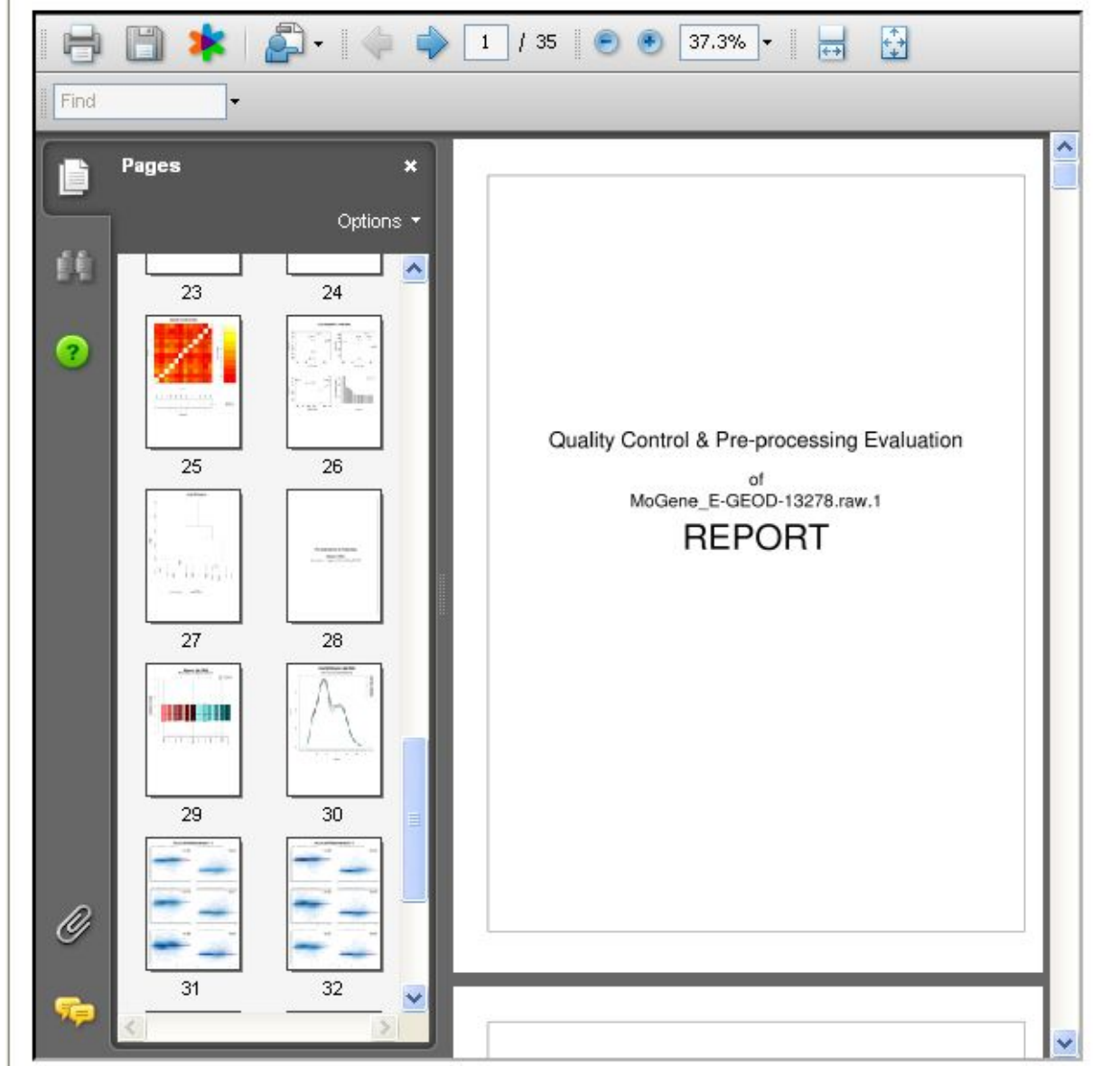
**Result files** (Right click on the following link(s) to save the corresponding file)

[Open log file](#) containing standard output, warning and error messages from the execution.  
You may also look over this text file on the following section: Output message (STDOUT & STDERR).

[Open PDF report file](#) contains all quality control images and results.  
You may also look over this PDF document in the frame below.

[Open zip file](#) with result tables and images (png format).

[Open pre-processed data](#) as a tab-delimited file.



The screenshot displays a PDF viewer interface. At the top, there is a toolbar with icons for printing, saving, and navigation, along with a zoom level of 37.3%. Below the toolbar is a search bar labeled 'Find'. The main content area shows a report titled "Quality Control & Pre-processing Evaluation of MoGene\_E-GEOD-13278.raw.1 REPORT". On the left side, there is a sidebar with a "Pages" panel showing thumbnails for pages 23 through 32. The thumbnails include various charts and graphs.

A second section of the result page shows the log file content. This information is important when you encountered a bug in the execution: you can report the bug in our [internal tracking system](#) or [by email](#). If you do so, please send us the log information by:

- either saving the log file on your computer (see previous links) and attach it to the ticket/e-mail
- or copy and paste the text from the screen.

### Output message (STDOUT & STDERR):

#### Standard output:

```
[1] "Parameters have been registered"
[1] "Zip file: MoGene_E-GEOD-13278.raw.1_2011-04-20_13-33_07.zip"
[1] "Raw data ready to be loaded in R"
[1] "Raw data have been loaded in R"
Script run using R version 2.13.0
registering new summary method 'pdnn'.
registering new pmcorrect method 'pdnn' and 'pdnnpredict'.
[1] "Libraries have been loaded"
[1] "Functions have been loaded"
current cdf environment loaded: MoGene-1_0-st-v1
The arrays are determined to contain perfect match probes only
[1] "Graphs ready to be computed"
[1] " plot degradation plot"
[1] " plot pos & neg control distribution"
[1] " plot control profiles and/or boxplots"
[1] " plot boxplot for raw intensities"
[1] " plot density histogram for raw intensities"
[1] " MA-plots for raw intensities"
[1] " plot array reference layout"
[1] " Pos/Neg COI"
[1] " Fit a probe level model (PLM) on the raw data"
[1] " 2D virtual images"
[1] " Complete set of 2D PLM images"
[1] " NUSE boxplot"
[1] " RLE boxplot"
[1] " Correlation plot of raw data"
[1] " Correlation plot of raw data"
[1] " Correlation plot of raw data"
```



## 3. ACKNOWLEDGMENTS

We would like to express our gratitude for using the open-access applications of ArrayAnalysis.org. This tutorial is derived from <http://www.arrayanalysis.org/> documentation originally written by Lars Eijssen and Anwasha Bohler.

The eNanoMapper project is funded by the European Union's Seventh Framework Program for research, technological development and demonstration (FP7-NMP-2013-SMALL-7) under grant agreement no. 604134.

## 4. REFERENCES

ArrayAnalysis homepage and web tools: <http://www.arrayanalysis.org/>

User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. Eijssen LM, Jaillard M, Adriaens ME, Gaj S, de Groot PJ, Müller M, Evelo CT. Nucleic Acids Res. 2013 Apr 24. PMID: [23620278](https://pubmed.ncbi.nlm.nih.gov/23620278/) doi: [10.1093/nar/gkt293](https://doi.org/10.1093/nar/gkt293)

## 5. KEYWORDS

Microarray data analysis  
Quality control and data pre-processing  
Affymetrix microarrays  
Systems biology  
Pathway and network analysis