CARMAweb users guide version 1.0.8

Johannes Rainer

July 4, 2006

Contents

Chapter 1

Introduction

The web application CARMAweb (Comprehensive R based Microarray Analysis web service) is an intuitive graphical user interface for the normalization and analysis of microarray data derived from the main current microarray platforms (Affymetrix GeneChips, two-color microarrays and ABI (Applied Biosystems) microarrays). The web application is build as a multi tier application based on the Java 2 Platform Enterprise Edition using the Struts framework [1] with Java Server Pages (JSP) and Servlets for the web tier and Enterprise Java Beans (EJB) for the business tier.

Because R [11] and BioConductor [6] are used as analytical backend a wide range of approved and state of the art normalization and analysis methods for all current microarray platforms are available in the web application. The cluster analysis module *GenesisWeb* uses facilities of the GenesisServer application server [14] to perform cluster analysis using a wide range of clustering algorithms.

A usermanagement guarantees password protected access to the users data and analysis results. Additionally all facilities of CARMAweb can be used without an user account, but in this case the user has to stay connected to the web application until all analysis have been completed; once logged out all data are deleted.

A typical analysis workflow in CARMAweb can be delineated as follows:

- Data upload:raw data files (e.g. Affymetrix **.CEL* files, GenePix **.gpr* files ...) can be uploaded to the web application using the *upload* in the *Data directory*. The users *Data directory* can be accessed using the appropriate link in the navigation tree (see figure 1.2).

- Preprocessing:using the *New Analysis* button in the button bar of the main page a new analysis can be defined. Depending on the platform used the user has to choose between a Affymetrix analysis, Two color microarray analysis, ABI microarray analysis or, if the analysis should be performed on already normalized data, Normalized data analysis (as the data is already normalized no preprocessing of the data will be performed).

- Replicate handling:the replicate handling step (see section 2.5) is part of the data preprocessing and allows to average the measured and normalized intensity values of replicated microarrays. This step can also be skipped if no replicates are available in the microarray experiment.

- Detection of differentially expressed genes:differentially expressed genes can be detected by using simple fold change cut–offs in the Fold change analysis, or using statistical tests.

- Cluster analysis:a cluster analysis can be performed independently of the previous steps. As input file any tabulator delimited text file containing expression or regulation values can be used. A new cluster analysis can be started using the *New Cluster analysis* button in the button bar of the main page, or by selecting a file for the cluster analysis in a result folder (see chapter 4).

- GO mapping analysis:the GO analysis can be started by the *New GO mapping* button in the button bar. As input any tabulator delimited text file containing in one column the EntrezGene (formerly LocusLink) identifiers for the genes of interest is accepted. GO analyses can be performed on files that are stored in the *Data directory*, or directly from within a R analysis result folder (see chapter 4).

All files (results) can be returned after each analysis from a analysis result folder (see chapter 4) to the users *Data directory* and therefore be used as input files for any further analysis. In general the microarray analysis can be split up into 3 main steps 1) Preprocessing, 2) detection of interesting (differentially expressed) genes and 3) *high level* analysis. This is a very general abstraction of the microarray analysis and usually researches will not need all of the steps mentioned above and will also have to change the order of the steps in this analytical pipeline. The design and modular concept of CARMAweb tries to model this analytical pipeline where the order of the modules can be changed and data and results can be transferred between its components. A description how this can be achieved in CARMAweb is described in figure 1.1.

The main web page of the CARMAweb web application can be split up into the three main parts button bar, navigation tree and center frame (see figure 1.2). The buttons from the button bar can be used to start a new microarray analysis (*New Analysis* see also figure 1.3), a new GO analysis (*New GO mapping*) or a new Cluster analysis (*New Cluster analysis*). The navigation tree in the left side of the web page can be used to navigate through the web application. It contains links to the users *Data directory* and to all analysis of the user. Each analysis is stored on the server until the user deletes them by clicking on the trashcan button next to the link in the navigation tree.

The users *Data directory* is secured by the usermanagement, that allows only the owner to access the directory. All files listed in the *Data directory* can be used to perform a new analysis. Data files can be uploaded into the users *Data directory* using the upload form in the *Data directory*. Additionally it is possible to fetch files from the cluster analysis module GenesisWebs workspace to the users *Data directory*, thus allowing to fetch back cluster analysis results (clusters exported to txt files) and to perform successive analyses on them. Raw data files can be normalized, normalized data can be used to perform detection of differentially expressed genes. All files from an analysis result folder (see chapter 4) can be copied to the *Data directory* and can then be used to perform a new analysis (e.g. cluster analysis or mapping to GO terms).

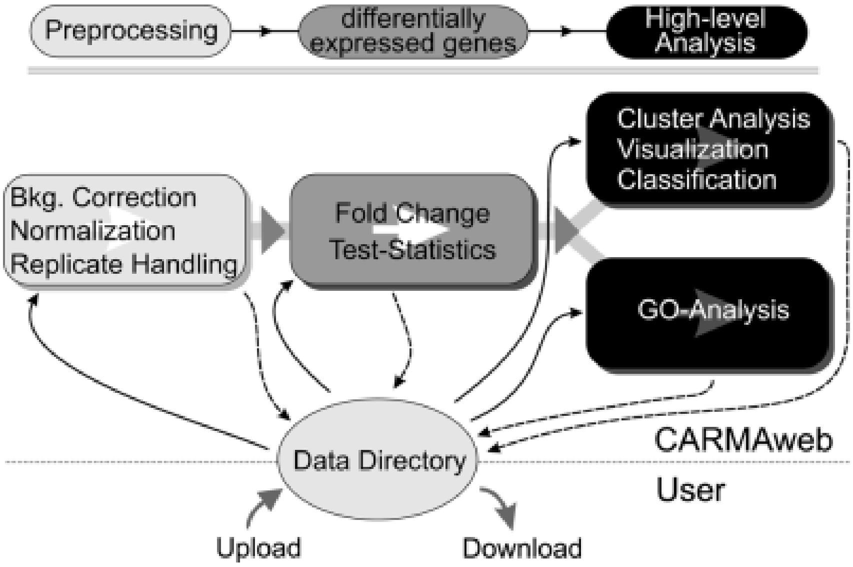
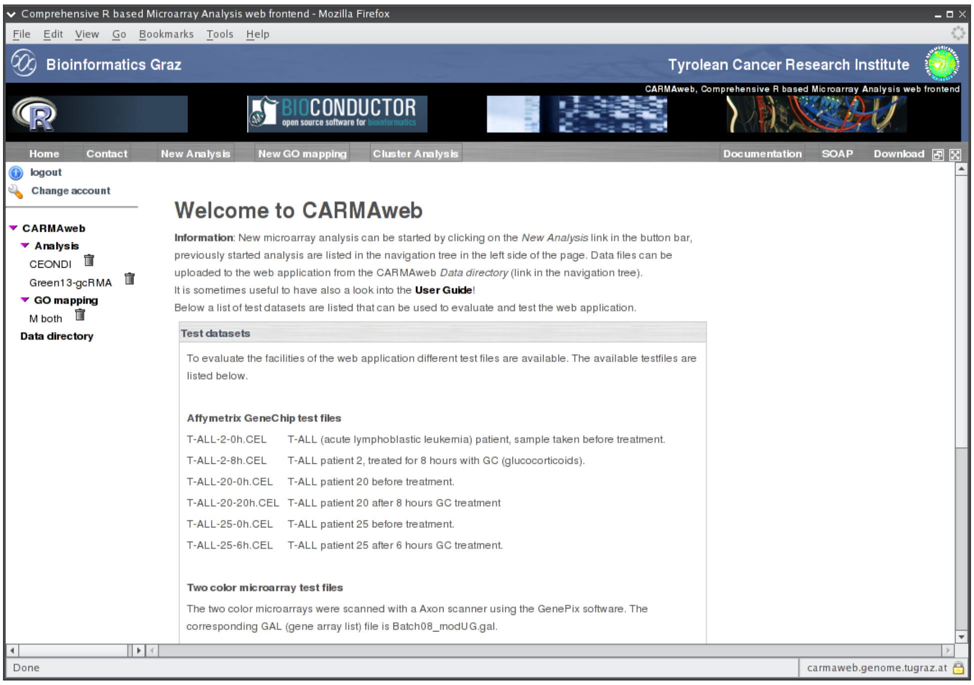


Figure 1.1: Microarray analysis workflow (pipeline) and how this workflow is modeled in CARMAweb. All the different modules of the web application can be used individually, or the can be combined to form a complete analytical pipeline. The central part of the workflow is the users `Data directory`, where analysis results can be returned and be used as input for any further analysis.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Welcome to CARMAweb

Information: New microarray analysis can be started by clicking on the *New Analysis* link in the button bar, previously started analysis are listed in the navigation tree in the left side of the page. Data files can be uploaded to the web application from the CARMAweb *Data directory* (link in the navigation tree).
It is sometimes useful to have also a look into the **User Guide**!
Below a list of test datasets are listed that can be used to evaluate and test the web application.

Test datasets

To evaluate the facilities of the web application different test files are available. The available testfiles are listed below.

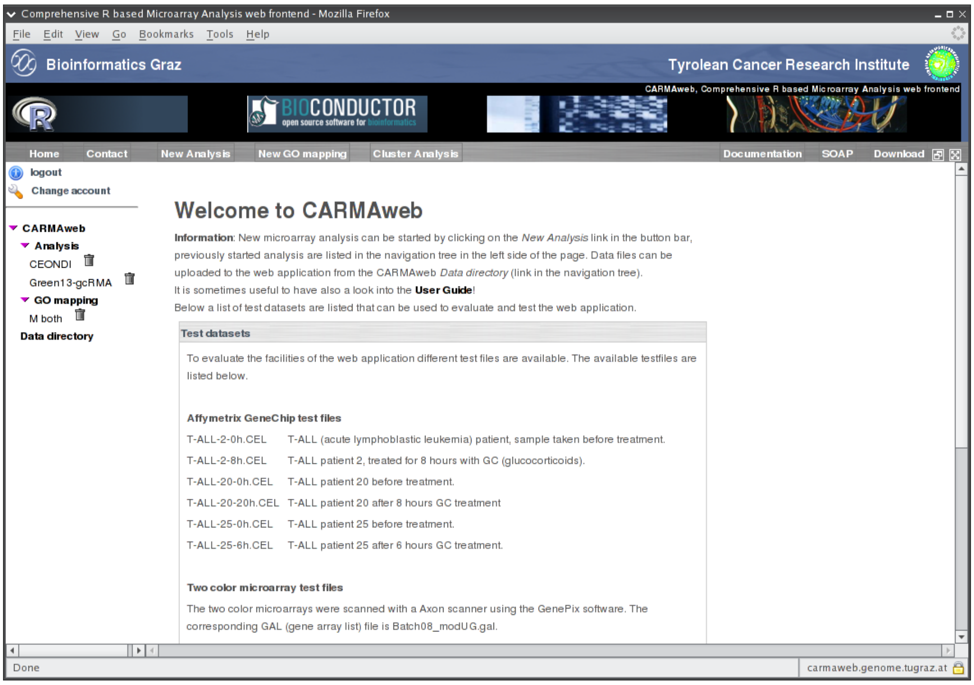
Affymetrix GeneChip test files

T-ALL-2-0h.CEL	T-ALL (acute lymphoblastic leukemia) patient, sample taken before treatment.
T-ALL-2-8h.CEL	T-ALL patient 2, treated for 8 hours with GC (glucocorticoids).
T-ALL-20-0h.CEL	T-ALL patient 20 before treatment.
T-ALL-20-20h.CEL	T-ALL patient 20 after 8 hours GC treatment
T-ALL-25-0h.CEL	T-ALL patient 25 before treatment.
T-ALL-25-6h.CEL	T-ALL patient 25 after 6 hours GC treatment.

Two color microarray test files

The two color microarrays were scanned with a Axon scanner using the GenePix software. The corresponding GAL (gene array list) file is Batch08_modUG.gal.

Figure 1.2: CARMAweb main page.


Welcome to CARMAweb

Information: New microarray analysis can be started by clicking on the *New Analysis* link in the button bar, previously started analysis are listed in the navigation tree in the left side of the page. Data files can be uploaded to the web application from the CARMAweb *Data directory* (link in the navigation tree).
It is sometimes useful to have also a look into the **User Guide**!
Below a list of test datasets are listed that can be used to evaluate and test the web application.

Test datasets

To evaluate the facilities of the web application different test files are available. The available testfiles are listed below.

Affymetrix GeneChip test files

T-ALL-2-0h.CEL T-ALL (acute lymphoblastic leukemia) patient, sample taken before treatment.
T-ALL-2-8h.CEL T-ALL patient 2, treated for 8 hours with GC (glucocorticoids).
T-ALL-20-0h.CEL T-ALL patient 20 before treatment.
T-ALL-20-20h.CEL T-ALL patient 20 after 8 hours GC treatment
T-ALL-25-0h.CEL T-ALL patient 25 before treatment.
T-ALL-25-6h.CEL T-ALL patient 25 after 6 hours GC treatment.

Two color microarray test files

The two color microarrays were scanned with a Axon scanner using the GenePix software. The corresponding GAL (gene array list) file is Batch08_modUG.gal.

Figure 1.3: New Analyses can be started from this page.

Chapter 2

Preprocessing

The preprocessing of microarray data includes the background correction, normalization of the expression values within each microarray and between the microarrays (to adjust the expression values accross microarrays). The aim of the data preprocessing step is to remove the technical variance and systematic errors without altering the biological variance within the data. Different adjustment methods have been developed in the last years for the different microarray platforms.

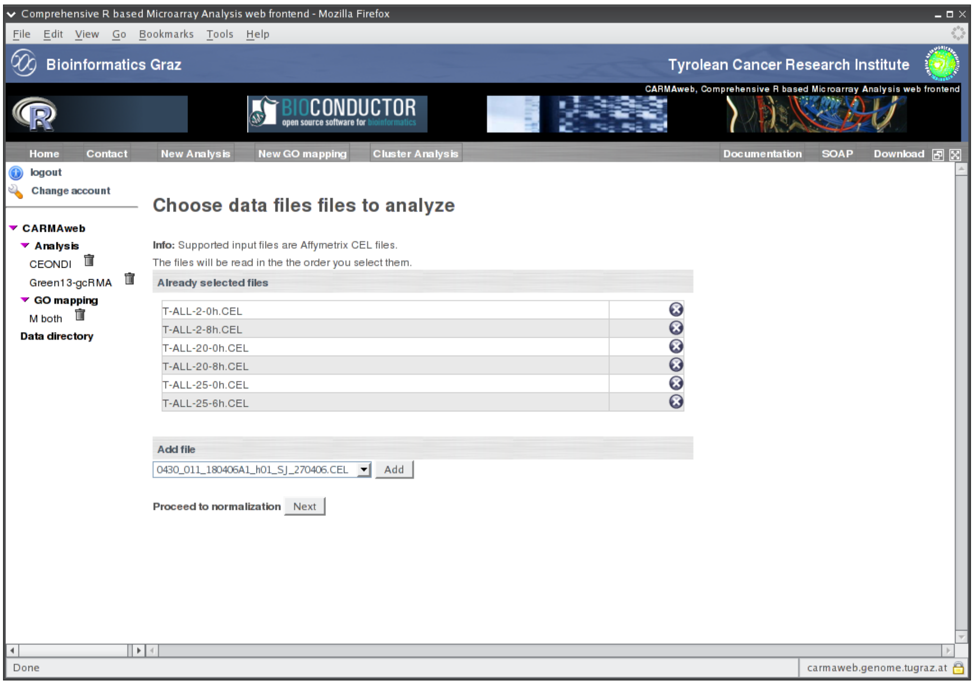
CARMAweb provides preprocessing methods for all different microarray platforms using facilities of the corresponding BioConductor packages. Affymetrix GeneChip data, spotted two color microarrays or ABI microarrays can be analyzed. The following sections give a brief introduction to the preprocessing steps of each of these platforms.

2.1 Preprocessing of Affymetrix GeneChip data

This section describes the pre–processing (background correction, normalization) for the Affymetrix GeneChip platform. CARMAweb uses the `affy` and `gcrma` packages from `BioConductor` for this purpose and allows therefore the usage of methods like the Affymetrix standard method MAS5 or more sophisticated methods like MBEI (model based expression index) [10], VSN (variance stabilizing normalization) [7], RMA (robust multi–array average) [9, 8] or GCRMA (robust multi–array average with correction for non specific binding according to the GC content of each probe sequence) to perform the preprocessing. Alternatively a custom normalization can be defined, where different methods can be used for the different steps of the Affymetrix GeneChip preprocessins (that are 1) background correction, 2) normalization, 3) correction for non specific binding and 4) summarization). A comparison of the different Affymetrix preprocessing methods is outlined in [3].

CARMAwebs preprocessing module for Affymetrix GeneChips supports *.CEL* files as input files. If you are planning to analyze Affymetrix GeneChips with a custom layout it is suggested that you contact the developer (email to *johannes.rainer@tcri.at*), as not all custom layout CDF environments are available from Bioconductor. The example analysis below has been performed using the Affymetrix test datasets available in CARMAweb. This test dataset consists of 6 Affymetrix GeneChips (hgu133plus2) on which samples from 3 T–ALL (acute lymphoblastic leukemia) patients before, and after treatment for 6 to 8 hours with GC (glucocorticoids), were hybridized. To analyze own data files it is required that the files are uploaded to CARMAweb (using the `Data directory`) before they can be processed.

The first step of the analysis consists in the selection of the files that should be preprocessed (see figure 2.1).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Choose data files files to analyze

Info: Supported input files are Affymetrix CEL files.
The files will be read in the the order you select them.

Already selected files

T-ALL-2-0h.CEL
T-ALL-2-8h.CEL
T-ALL-20-0h.CEL
T-ALL-20-8h.CEL
T-ALL-25-0h.CEL
T-ALL-25-6h.CEL

Add file

0430_011_180406A1_h01_SJ_270406.CEL [Add]

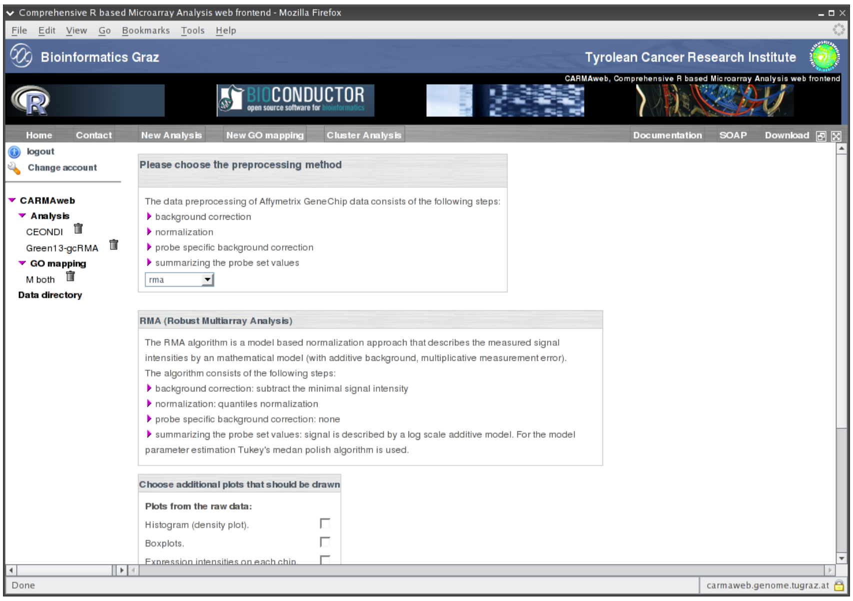
Proceed to normalization [Next]

Figure 2.1: Selection for the Affymetrix GeneChip raw data files that should be analyzed.

Next the preprocessing method can be selected (figure 2.2). The Affymetrix preprocessing consists of the steps:

- Background correction
- Normalization
- Probe specific background correction
- Summarization of the probe level data

CARMAweb provides various preprocessing methods (like *rma*, *gcrma*...) and allows also to perform a *custom* preprocessing by selecting different algorithms for the various steps of the preprocessing. For the preprocessing of large Affymetrix GeneChip experiments including many GeneChips the *justRMA* or *justGCRMA* methods are suggested, as these versions of the *rma* and *gcrma* algorithm are faster and use less memory.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Please choose the preprocessing method

The data preprocessing of Affymetrix GeneChip data consists of the following steps:

- background correction
- normalization
- probe specific background correction
- summarizing the probe set values

rma

RMA (Robust Multiarray Analysis)

The RMA algorithm is a model based normalization approach that describes the measured signal intensities by an mathematical model (with additive background, multiplicative measurement error).
The algorithm consists of the following steps:

- background correction: subtract the minimal signal intensity
- normalization: quantiles normalization
- probe specific background correction: none
- summarizing the probe set values: signal is described by a log scale additive model. For the model parameter estimation Tukey's medan polish algorithm is used.

Choose additional plots that should be drawn

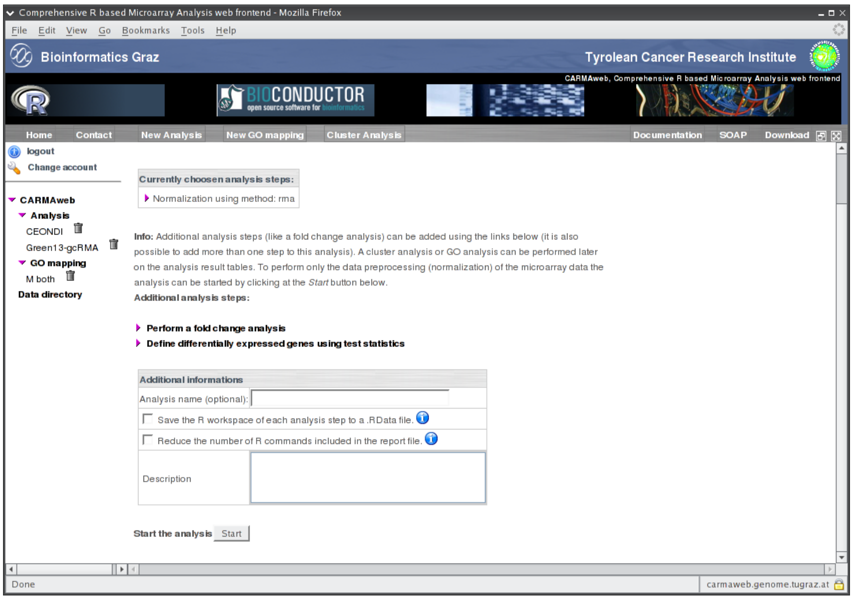
Plots from the raw data:

Histogram (density plot). □

Boxplots. □

Expression intensities on each chip □

Figure 2.2: Affymetrix preprocessing.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Currently choosen analysis steps:

- Normalization using method: rma

Info: Additional analysis steps (like a fold change analysis) can be added using the links below (it is also possible to add more than one step to this analysis). A cluster analysis or GO analysis can be performed later on the analysis result tables. To perform only the data preprocessing (normalization) of the microarray data the analysis can be started by clicking at the *Start* button below.

Additional analysis steps:

- **Perform a fold change analysis**
- **Define differentially expressed genes using test statistics**

Additional informations

Analysis name (optional):

□ Save the R workspace of each analysis step to a .RData file.

□ Reduce the number of R commands included in the report file.

Description

Start the analysis Start

Figure 2.3: Main analysis page for Affymetrix GeneChips. Additional analysis steps can be added by clicking on the appropriate links. The analysis can be started by clicking on the *Next* button.

By clicking on the *Analysis* button the user gets forwarded to the analysis page (figure 2.3), where additional analysis steps, like detection for differentially expressed genes, can be added to the analysis (the detection of differentially expressed genes for this example using test statistics is described in section 3.2).
The option *Save the R workspace of each analysis step to a .RData file* (see figure 2.3) makes the whole R work space (containing all the objects created and used for the analysis) available for download, so that the user can reload it on his local computer back into R (using R's `load` function).
The analysis can be started by clicking on the *Next* button on the page (see figure 2.3).

2.2 Preprocessing of two color microarray data

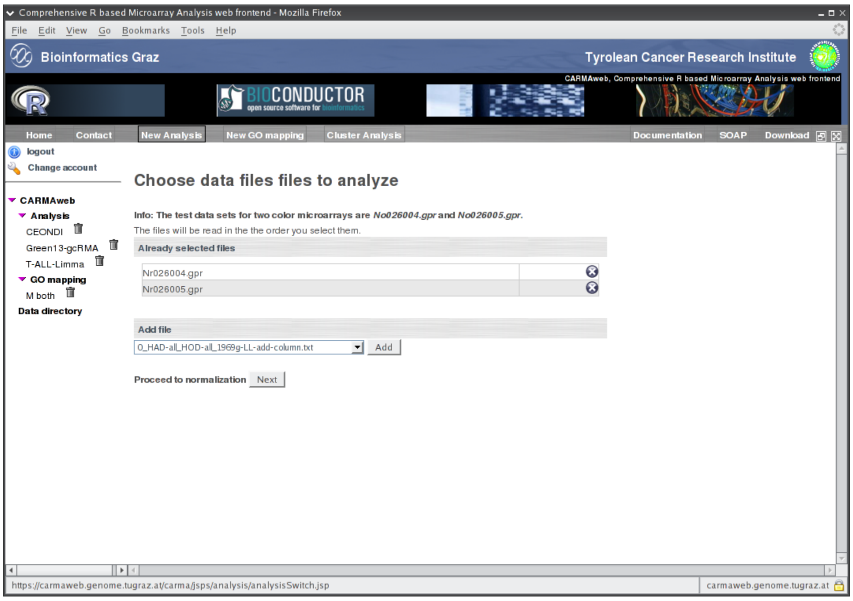
The preprocessing of two color microarrays consists of the steps 1) background correction, 2) within microarray normalization and (optional) 3) between microarray normalization. The background correction can be performed in CARMAweb by subtracting the background from the foreground signal (method `subtract`) or by using more advanced methods like the `minimum` method (any intensity wich is zero or negative after correction is set to half the minimum of the positive corrected intensities), the `moving minimum` method (background estimates are replaced with the minimum of the backgrounds of the spot and its eight neighbors) or the method described in [5] (`edwards`).

For the within microarray normalization methods like the `median` normalization as well as the `loess` or `print tip loess` normalization, the `robust spline` normalization (normalizes using robustly fitted regression splines and empirical Bayes shrinkage) or the variance stabilizing normalization (`vsn`) [7] (which combines the within and between array normalization) are provided, whereas for the between microarray normalization the median scaling or the quantile method are available. Most of these preprocessing methods are outlined in [12]. The preprocessing of two color microarrays is carried out in CARMAweb using functions from Bioconductors limma and vsn packages. As input files raw data files from a wide range of microarray image analysis softwares are supported (*Agilent Feature Extraction*, *ArrayVision*, *BlueFuse*, *GenePix*, *ImaGene*, *QuantArray*, *SPOT*, or raw data files from the Stanford Microarray Database (*SMD*)). The different plots are created using functions from the `maDB` package which provides also the functionality for the replicate handling step. Technically replicated arrays or replicated spots on each array can be merged by calculating the median or mean expression value across the replicates.

Genes flagged as bad spots by the scanning software can be excluded from the analysis (for Genepix users: choose *give all flagged spots a weight of 0* where *How should spots flagged by the software be handled?* is asked).

The following analysis was performed using the two color microarray test files available in the web application. The files were scanned in a Axon scanner using the GenePix software (version 5). The annotation for each spot is available in the corresponding *Batch08_modUG.gal* file. GenePix users actually do not have to provide the annotation using a GAL file, as the annotation is by default included in the *gpr* files itself. GenePix users can also choose which columns of the input files (gpr files) should be used as the red and green foreground respectively background signal channels.

For the two-color microarray analysis the two test files available in CARMAweb were used (see figure 2.4).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
https://carmaweb.genome.tugraz.at/carma/jsps/analysis/analysisSwitch.jsp
carmaweb.genome.tugraz.at

Choose data files files to analyze

Info: The test data sets for two color microarrays are *No026004.gpr* and *No026005.gpr*.

The files will be read in the the order you select them.

Already selected files

Nr026004.gpr

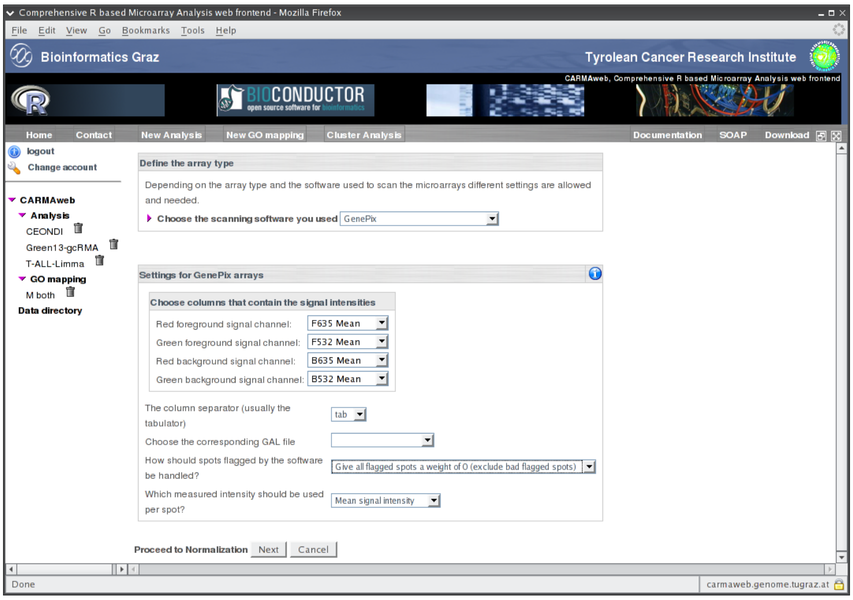
Nr026005.gpr

Add file

0_HAD-all_HOD-all_1969g-LL-add-column.txt Add

Proceed to normalization Next

Figure 2.4: Like in the Affymetrix analysis the data files containing the raw data files can be added to the analysis using the page shown above.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Define the array type

Depending on the array type and the software used to scan the microarrays different settings are allowed and needed.

Choose the scanning software you used GenePix

Settings for GenePix arrays

Choose columns that contain the signal intensities

Red foreground signal channel: F635 Mean

Green foreground signal channel: F532 Mean

Red background signal channel: B635 Mean

Green background signal channel: B532 Mean

The column separator (usually the tabulator) tab

Choose the corresponding GAL file

How should spots flagged by the software be handled? Give all flagged spots a weight of 0 (exclude bad flagged spots)

Which measured intensity should be used per spot? Mean signal intensity

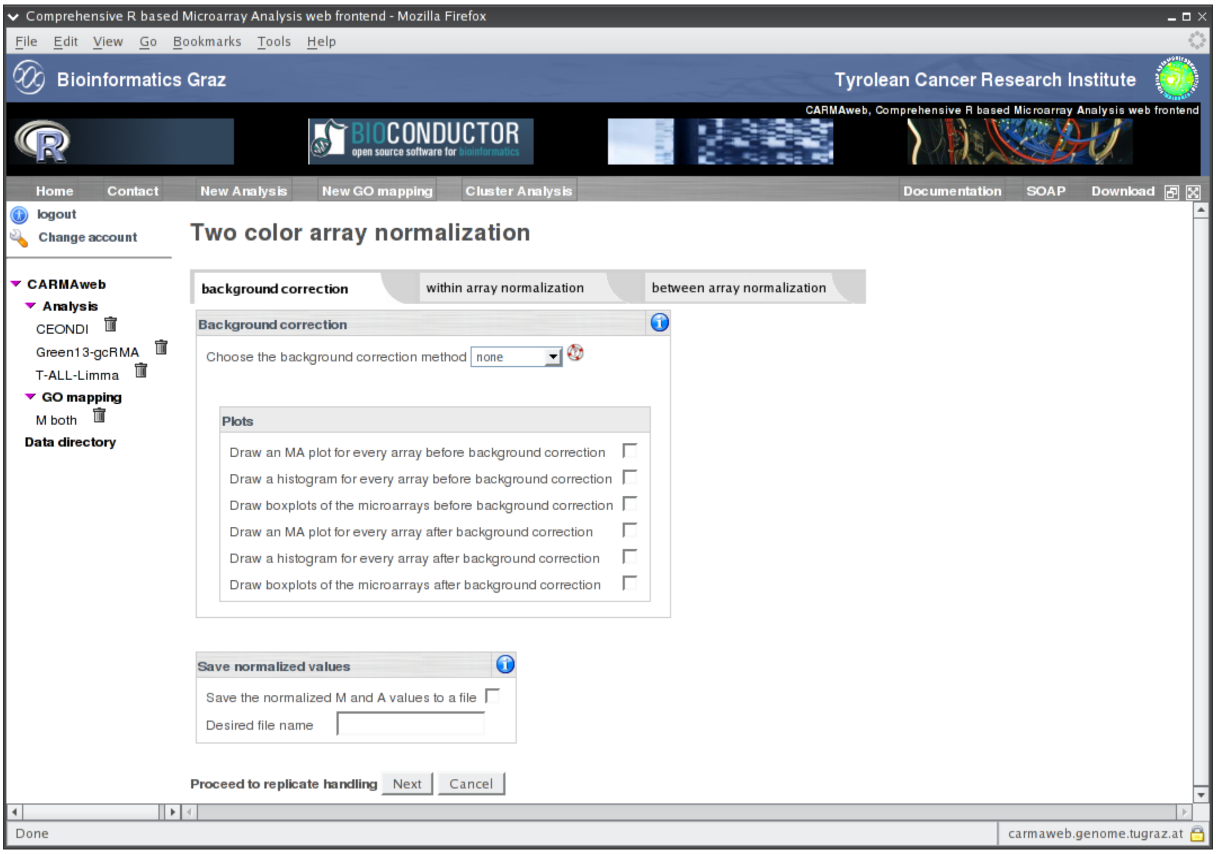
Proceed to Normalization Next Cancel

Figure 2.5: The software with which the microarrays were scanned has to be specified to allow the analysis of the raw data files. In the example above the GenePix software was used. Different parameters specific to the used scanning software can be defined also on this page..

The user has to specify the software with which the microarrays were scanned (figure 2.5) and can also select different parameters specific for the selected scanning software. In our example we selected to exclude all spots that were flagged as *bad spots* by the GenePix scanning sofware (option *Give all flagged spots a weight of 0*). Additionally is possible to choose the data columns from the raw data files that should be used to read the signal intensities for the red and green foreground respectively background (e.g. by choosing *F635 Median* for the red foreground channel the median intensity per spot would be used instead of the mean instensity).

The preprocessing for two color microarrays consists of the three steps Background correction, within array normalization and between array normalization. The settings for each one of these steps can be defined on the different pages *background correction*, *within array normalization* and *between array normalization*. It is possible to switch between these pages by clicking on the appropriate tab (see figure 2.6, tabs *background correction*, *within array normalization* and *between array normalization*).

The background correction method as well as various plots that should be generated before and after background correction can be selected on the page shown in figure 2.6.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
background correction
within array normalization
between array normalization
Next
Cancel
Done
carmaweb.genome.tugraz.at

Two color array normalization

Background correction

Choose the background correction method none

Plots

Draw an MA plot for every array before background correction ☐
Draw a histogram for every array before background correction ☐
Draw boxplots of the microarrays before background correction ☐
Draw an MA plot for every array after background correction ☐
Draw a histogram for every array after background correction ☐
Draw boxplots of the microarrays after background correction ☐

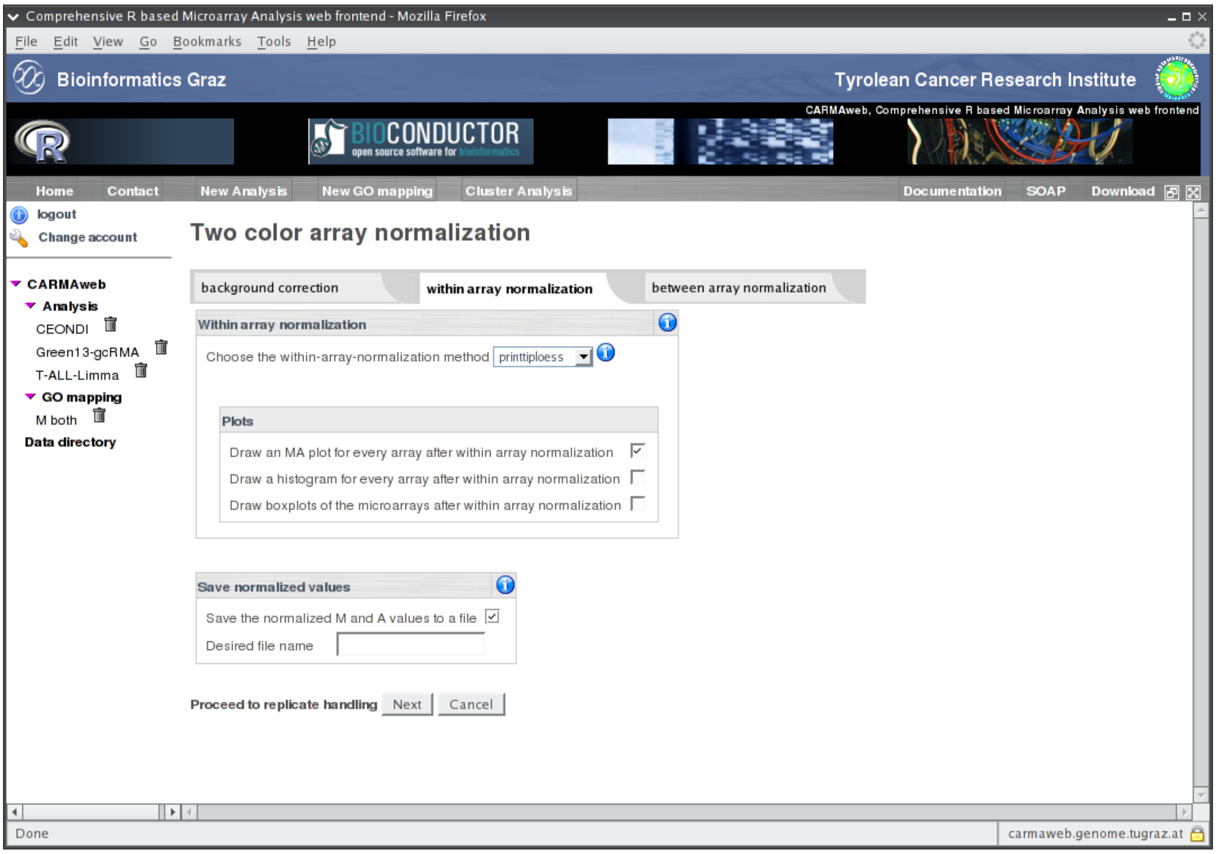
Save normalized values

Save the normalized M and A values to a file ☐
Desired file name

Proceed to replicate handling

Figure 2.6: Background correction of two color microarrays.

The within array normalization corrects for the dye bias and other systematic errors within each microarray. The default method in CARMAweb is the print tip loess normalization, which performs a normalization for each print tip group separately to account also for the print tip effect. The within array normalization method can be selected on the *within array normalization* tab of the two color microarray normalization page (see figure 2.7).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR
open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
background correction
within array normalization
between array normalization
Proceed to replicate handling
Next
Cancel
Done
carmaweb.genome.tugraz.at

Two color array normalization

Within array normalization

Choose the within-array-normalization method printtiploess

Plots

Draw an MA plot for every array after within array normalization ☑

Draw a histogram for every array after within array normalization □

Draw boxplots of the microarrays after within array normalization □

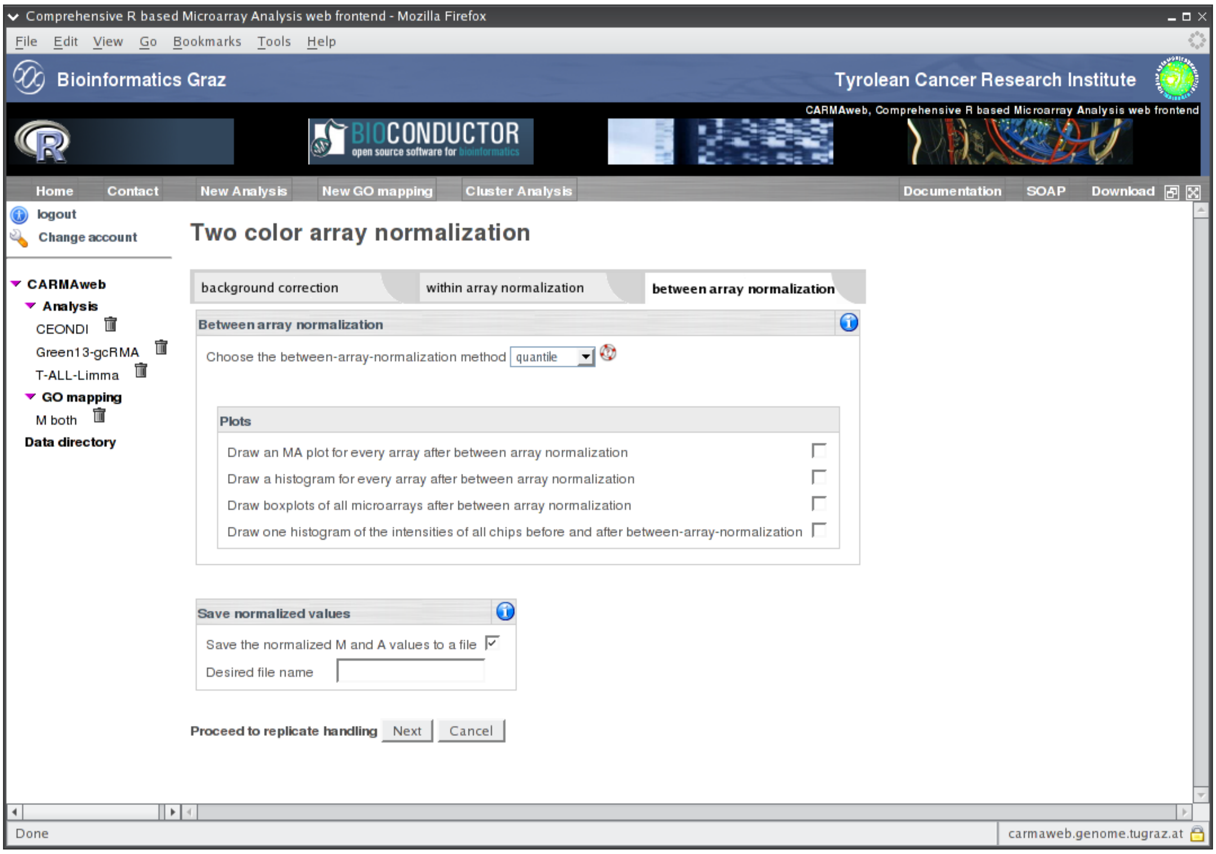
Save normalized values

Save the normalized M and A values to a file ☑

Desired file name

Figure 2.7: Within array normalization of two color microarrays.

The between array normalization (figure 2.8) adjusts the expression values for each spot across the microarrays of a microarray experiment. Available methods are a simple scaling method or the quantile mothod (known from the Affymetrix platform).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Two color array normalization

background correction | within array normalization | between array normalization

Between array normalization

Choose the between-array-normalization method quantile

Plots

Draw an MA plot for every array after between array normalization □

Draw a histogram for every array after between array normalization □

Draw boxplots of all microarrays after between array normalization □

Draw one histogram of the intensities of all chips before and after between-array-normalization □

Save normalized values

Save the normalized M and A values to a file ☑

Desired file name

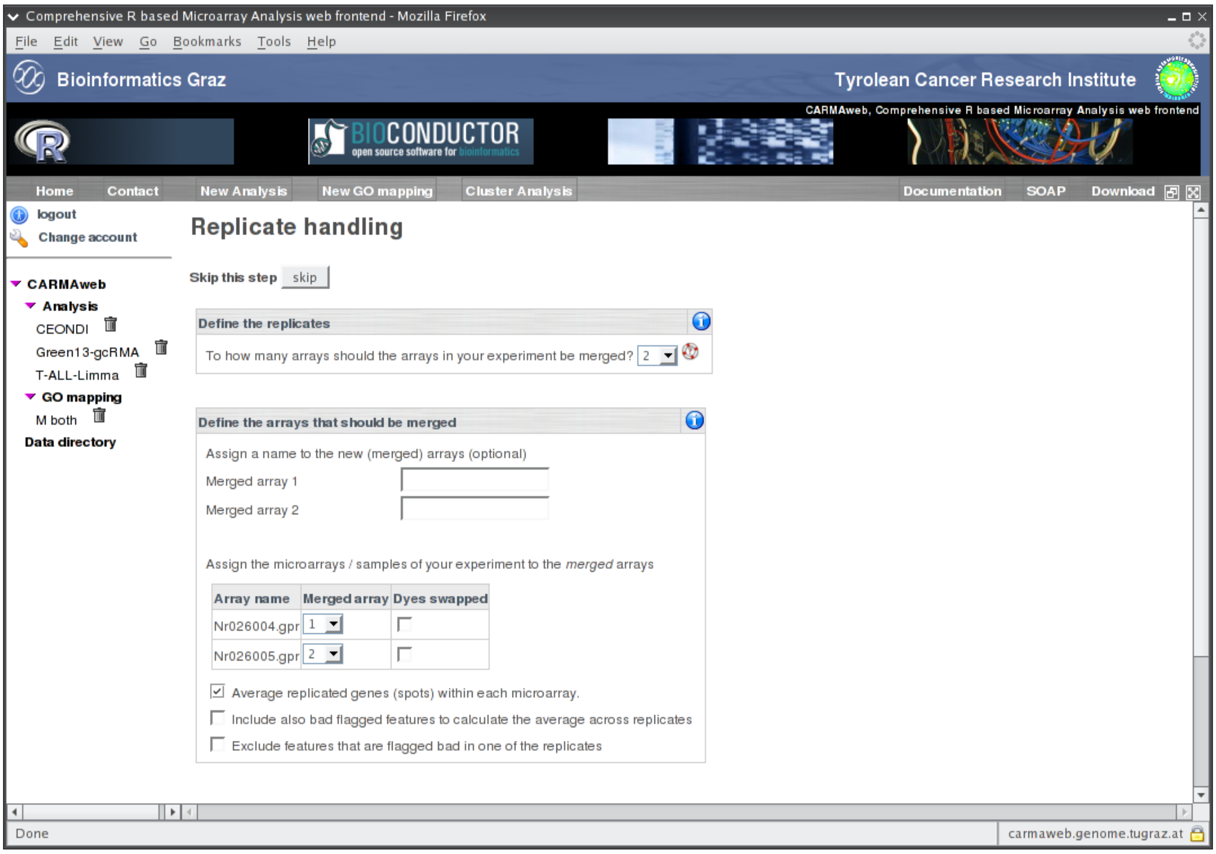
Proceed to replicate handling Next Cancel

Figure 2.8: Between array normalization of two color microarrays.

After the definition of the preprocessing it is possible to perform a simple replicate handling (figure 2.9), which allows to average the expression values accross replicated microarrays (e.g. average for each spot the signal intensities from the red signal channels of different microarrays). Different options that are available are:

- *Average replicated genes (spots) within each microarray* which averages the expression values from replicated spots within each microarray (given that these spots have the same gene identifier).
- *Include also bad flagged features to calculate the average across replicates*: will discard any information from the scanner software about the spot quality and use all spots of replicated microarrays to calculate the average expression value. This option is especially useful in experiment where e.g. only a *dye swap normalization* should be performed.
- *Exclude features that are flagged bad in one of the replicates*: expression measures of features that have one spot in the replicated microarrays that is flagged bad will not be averaged. For these features a NA value will be returned by the function (and therefore these features will be excluded from all further analyses).

In our example all replicated spots on each array should be averaged (option *Average replicated genes (spots) within each microarray*). The number of arrays to which the microarrays should be merged was therefore set to 2, and each microarray was assigned to one of the so called *merged arrays*. If replicated arrays are available in the microarray experiment (technical replicates), these arrays can be merged, by assigning the replicated microarrays to one *merged array*.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Replicate handling

Skip this step skip

Define the replicates

To how many arrays should the arrays in your experiment be merged? 2

Define the arrays that should be merged

Assign a name to the new (merged) arrays (optional)

Merged array 1

Merged array 2

Assign the microarrays / samples of your experiment to the *merged* arrays

Array name	Merged array	Dyes swapped
Nr026004.gpr	1	□
Nr026005.gpr	2	□

☑ Average replicated genes (spots) within each microarray.

□ Include also bad flagged features to calculate the average across replicates

□ Exclude features that are flagged bad in one of the replicates

Figure 2.9: Replicate handling. On this page replicated spots within a microarray or technical replicated microarrays can be averaged (merged). The replicate handling step can be skipped by clicking on the *skip* button.

After the replicate handling step the analysis can be started or additional steps can be added to the experiment (like the detection of differentially expressed genes). The detection of differentially expressed genes using fold change analysis for this example is described in chapter 3.1.

2.3 ABI microarray normalization

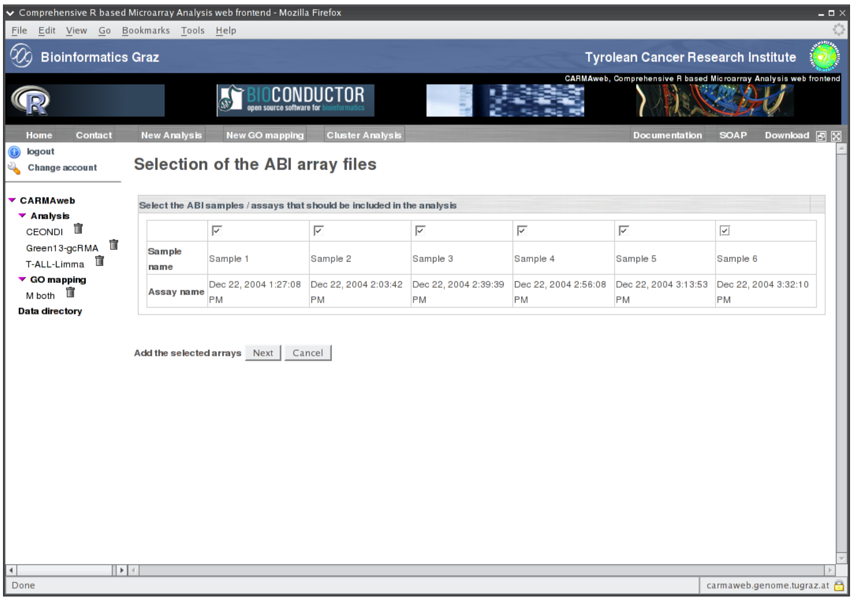
CARMAweb can read and process ABI microarray data exported as tabulator delimited txt files by the ABI scanning software. Required columns that have to be exported are:

- SIGNAL
- ASSAY_NORMALIZED_SIGNAL
- FLAGS
- PROBE_ID

- GENE_ID
- ASSAY_NAME
- SAMPLE_NAME

To perform a microarray analysis on ABI microarray data first the data file(s) have to uploaded using the *upload* link in the users CARMAweb *Data directory*. After uploading the file(s) select *New Analysis* on the button bar and choose *Perform an ABI microarray analysis*. Next the first data file has to be selected that contains the data from the ABI microarrays. Upon selecting the file, CARMAweb tries to read the sample and assay informations from this file and proceeds to the page shown in figure 2.10. The test file used in the following example (*ABIdata-mod2.txt*[1]) contains the data from 6 ABI microarrays (samples are labeled with *Sample 1 - 6*).

The samples (assays, microarrays) that should be included in the analysis can be selected using the page shown in figure 2.10. It is possible to load the microarray data of ABI microarrays from more than one text file. In the web page shown in figure 2.11 the microarrays respectiveley samples are listed that have already been added to the microarray analysis. Additionally other microarrays can be added to the analysis by selecting another file containing the corresponding data from other ABI microarrays.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Selection of the ABI array files

Select the ABI samples / assays that should be included in the analysis

	☑	☑	☑	☑	☑	☑
Sample name	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Assay name	Dec 22, 2004 1:27:08 PM	Dec 22, 2004 2:03:42 PM	Dec 22, 2004 2:39:39 PM	Dec 22, 2004 2:56:08 PM	Dec 22, 2004 3:13:53 PM	Dec 22, 2004 3:32:10 PM

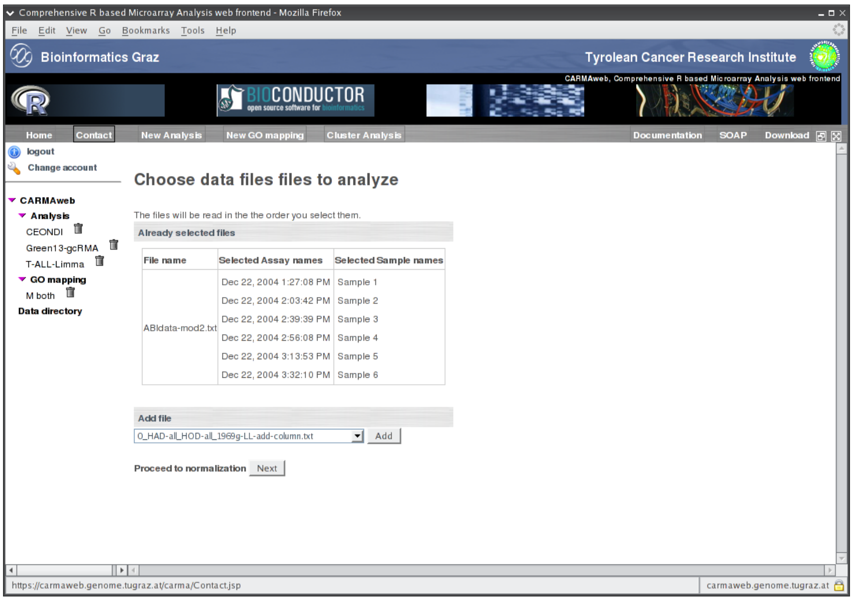
Add the selected arrays Next Cancel

Figure 2.10: Selection of the ABI microarrays for the analysis. The microarrays / samples that should be analyzed can be selected by checking the checkbox above the corresponding sample.

[1] As the data set contains unpublished data the gene and probe names were permutated

After adding all ABI microarrays to the analysis the pre–processing of the raw data values can be performed. The different parameters can be set in the according *ABI normalization* page (figure 2.12). As raw data signal either the *Assay normalized signal* or the raw signal can be used. The Assay normalized signal are an already background corrected expression values, that were global median adjusted accross all microarrays using the ABI built–in algorithm. The raw signal consists of background corrected raw expression values. These values have to be adjusted accross all microarrays in the experiment to allow comparisons between samples. The default normalization algorithm is the *quantile* method that is also used to adjust Affymetrix GeneChip data. Additionally different plots (histograms) can be drawn from the raw and from the normalized data and the normalized expression values can be saved to a text file. Probes can also be filtered using the FLAG values from the ABI software. Usually probes that have a FLAG value below 1000 are marked as *bad spots* and are excluded from all further analysis steps. The following replicate handling step (see also chapter 2.5 of the CARMAweb user guide) allows to average replicated microarrays and / or replicated probes on each microarray. The replicate handling can also be skipped if no replicated microarrays are present in the microarray experiment.

After selecting the normalization and replicate handling parameters additional analysis steps can be added to the analysis using the ABI analysis page shown in figure 2.13. Such steps can be the detection of differentially expressed genes using simple fold change analysis, or using statistical tests. The fold change analysis is described in chapter 3.1 and the detection of differentially expressed genes using test statistics in chapter 3.2 of the CARMAweb user guide.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
Documentation
SOAP
Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
https://carmaweb.genome.tugraz.at/carma/Contact.jsp
carmaweb.genome.tugraz.at

Choose data files files to analyze

The files will be read in the the order you select them.

Already selected files

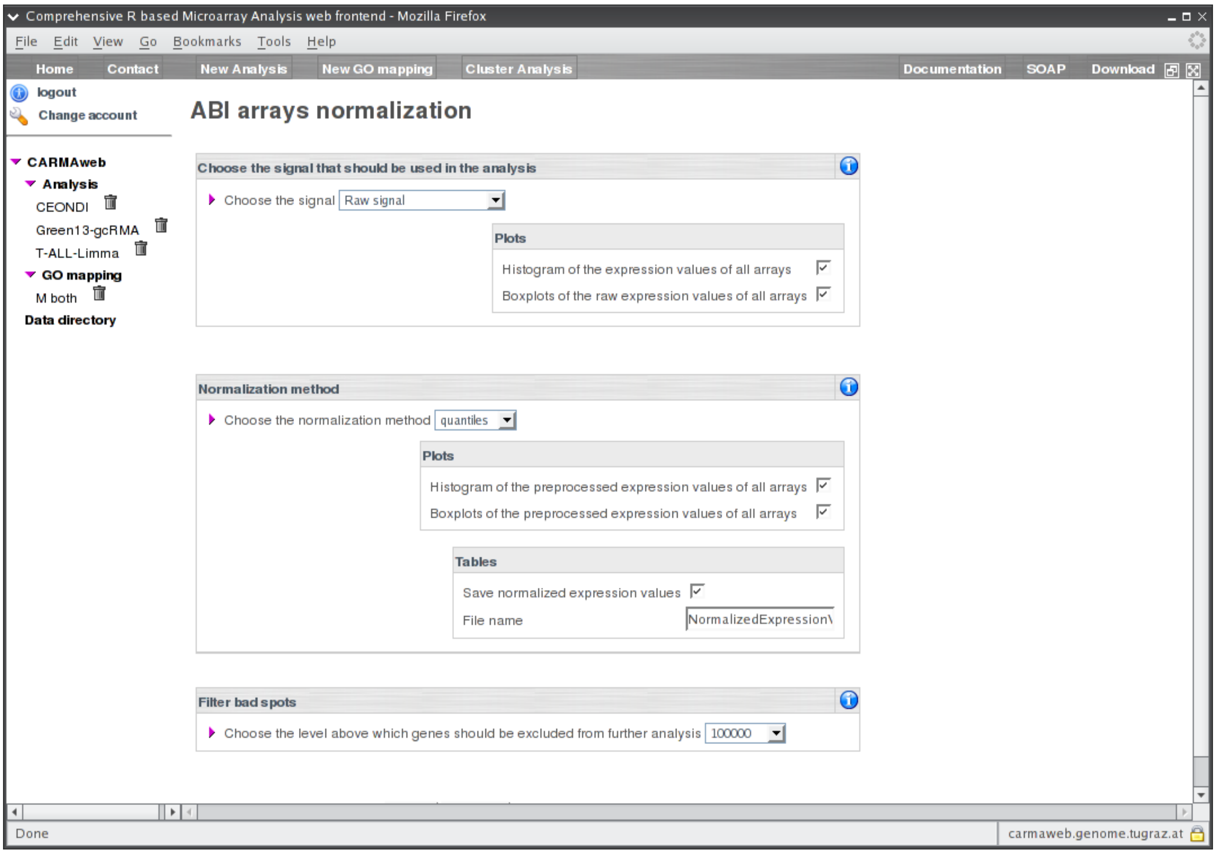
File name	Selected Assay names	Selected Sample names
ABIdata-mod2.txt	Dec 22, 2004 1:27:08 PM	Sample 1
	Dec 22, 2004 2:03:42 PM	Sample 2
	Dec 22, 2004 2:39:39 PM	Sample 3
	Dec 22, 2004 2:56:08 PM	Sample 4
	Dec 22, 2004 3:13:53 PM	Sample 5
	Dec 22, 2004 3:32:10 PM	Sample 6

Add file

0_HAD-all_HOD-all_1969g-LL-add-column.txt Add

Proceed to normalization Next

Figure 2.11: The ABI microarrays / samples that have been added to the analysis are listed. Data from other microarrays might be added to the analysis by adding a new text file that contains the data for the corresponding microarrays.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

ABI arrays normalization

Choose the signal that should be used in the analysis

Choose the signal: Raw signal

Plots

Histogram of the expression values of all arrays ☑

Boxplots of the raw expression values of all arrays ☑

Normalization method

Choose the normalization method: quantiles

Plots

Histogram of the preprocessed expression values of all arrays ☑

Boxplots of the preprocessed expression values of all arrays ☑

Tables

Save normalized expression values ☑

File name: NormalizedExpression\

Filter bad spots

Choose the level above which genes should be excluded from further analysis: 100000

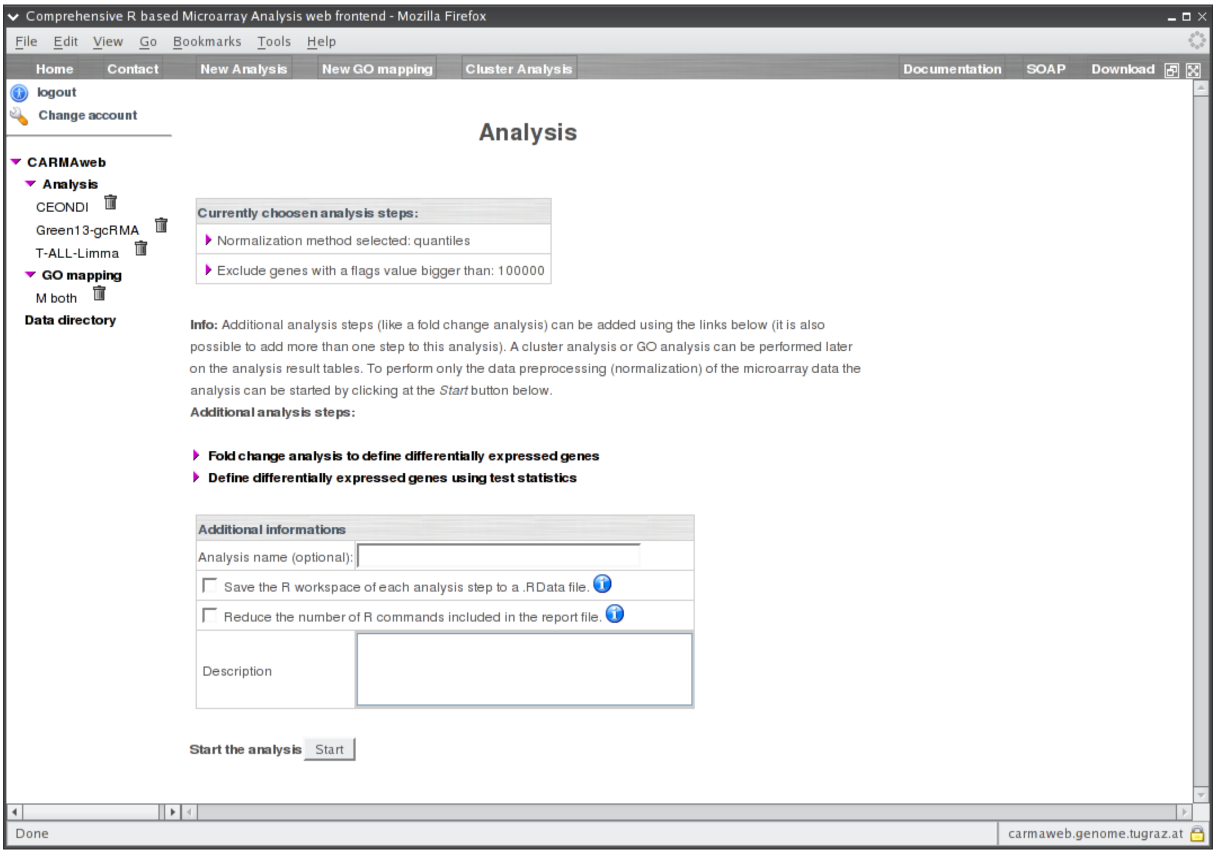
Figure 2.12: Normalization settings for ABI microarray data.

After the analysis with CARMAweb the sets of interesting genes can also be analyzed and annotated using the free ABI web applicaton `Panther` (https://panther.appliedbiosystems.com/).

2.4 Analysis using already normalized data

The detection of differentially expressed genes can also be performed directly on already normalized expression values (for example if the raw data was normalized in a previous analysis and the table with the normalized expression values of all genes has been copied from the analysis result folder (see chapter 4) to the users *Data directory*). This can be achieved by selecting *Perform an analysis on already normalized data* after clicking the *New Analysis* button. On the next page (see figure 2.14) the file containing the normalized expression values can be selected. The user has to choose the columns that contain the normalized expression values (all other columns should not be selected! they will be used as annotation columns), additionally it is possible to define the decimal point (either . or ,) that is used or to set the data to be log2 transformed. All columns that are not selected are used as *annotation* columns!

As input file any tabulator delimited txt file containing numerical and text data is supported. This allows to analyze not only microarray data (normlaized microarray data), but to analyze any data with the methods and algorithms


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Analysis

Currently choosen analysis steps:

- Normalization method selected: quantiles
- Exclude genes with a flags value bigger than: 100000

Info: Additional analysis steps (like a fold change analysis) can be added using the links below (it is also possible to add more than one step to this analysis). A cluster analysis or GO analysis can be performed later on the analysis result tables. To perform only the data preprocessing (normalization) of the microarray data the analysis can be started by clicking at the *Start* button below.

Additional analysis steps:

- **Fold change analysis to define differentially expressed genes**
- **Define differentially expressed genes using test statistics**

Additional informations

Analysis name (optional):

□ Save the R workspace of each analysis step to a .RData file.

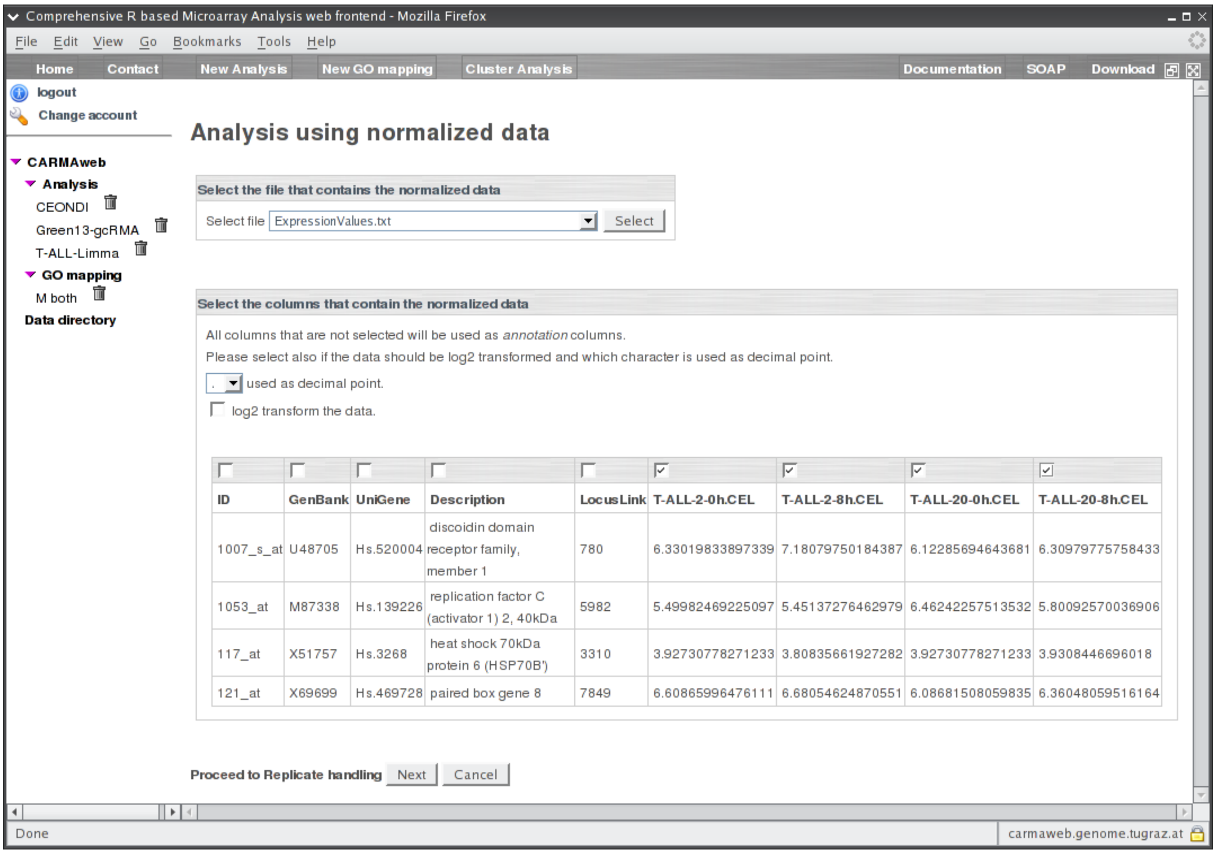
□ Reduce the number of R commands included in the report file.

Description

Start the analysis [Start]

Figure 2.13: ABI microarray analysis page. Additional analysis steps (like a simple fold change analysis or test statistics to determine differentially expressed genes) can be added to the analysis, or the analysis can be started by clicking on the *Start* button.

provided by CARMAweb.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Analysis using normalized data

Select the file that contains the normalized data

Select file ExpressionValues.txt Select

Select the columns that contain the normalized data

All columns that are not selected will be used as *annotation* columns.

Please select also if the data should be log2 transformed and which character is used as decimal point.

. used as decimal point.

□ log2 transform the data.

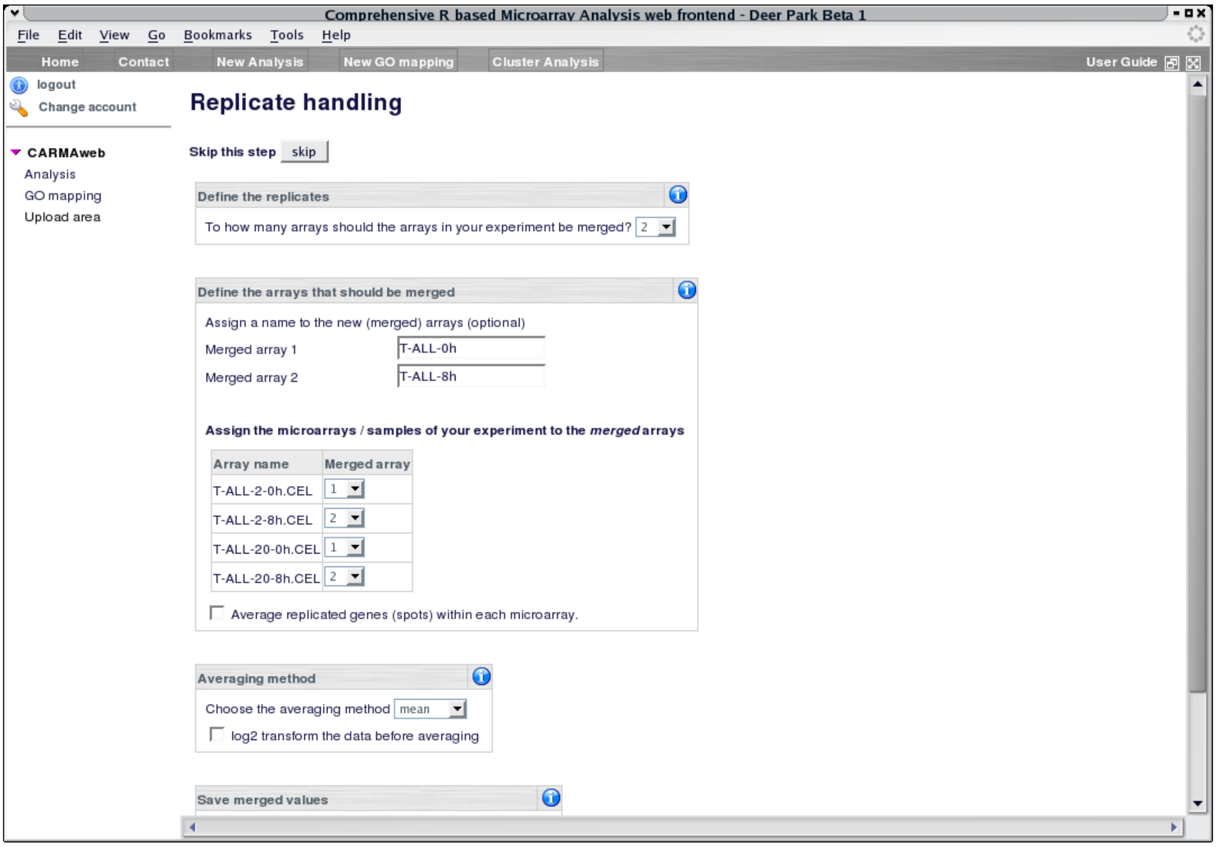
□	□	□	□	□	☑	☑	☑	☑
ID	**GenBank**	**UniGene**	**Description**	**LocusLink**	**T-ALL-2-0h.CEL**	**T-ALL-2-8h.CEL**	**T-ALL-20-0h.CEL**	**T-ALL-20-8h.CEL**
1007_s_at	U48705	Hs.520004	discoidin domain receptor family, member 1	780	6.33019833897339	7.18079750184387	6.12285694643681	6.30979775758433
1053_at	M87338	Hs.139226	replication factor C (activator 1) 2, 40kDa	5982	5.49982469225097	5.45137276462979	6.46242257513532	5.80092570036906
117_at	X51757	Hs.3268	heat shock 70kDa protein 6 (HSP70B')	3310	3.92730778271233	3.80835661927282	3.92730778271233	3.9308446696018
121_at	X69699	Hs.469728	paired box gene 8	7849	6.60865996476111	6.68054624870551	6.08681508059835	6.36048059516164

Proceed to Replicate handling Next Cancel

Figure 2.14: Analysis using already normalized data. After selecting the file, the first 4 rows are displayed, so that the user can select the columns containing the normalized expression values.

2.5 Replicate handling

The replicate handling step allows to average the normalized intensity (expression) values of replicated microarrays within a microarray experiment. Additionally replicated spots (genes or probe sets for Affymetrix GeneChips) can be averaged within each microarray. All spots that are flagged by the scanning software as bad spots are excluded from the replicate handling step (provided that this information was available in the input data and that the user has selected the appropriate settings to read this information from the input files (for Genepix users: choose *give all flagged spots a weight of 0* where *How should spots flagged by the software be handled?* is asked; ABI users can use the flags from the input file)).


Comprehensive R based Microarray Analysis web frontend - Deer Park Beta 1
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis User Guide
logout
Change account
CARMAweb
Analysis
GO mapping
Upload area

Replicate handling

Skip this step skip

Define the replicates

To how many arrays should the arrays in your experiment be merged? 2

Define the arrays that should be merged

Assign a name to the new (merged) arrays (optional)

Merged array 1 T-ALL-0h

Merged array 2 T-ALL-8h

Assign the microarrays / samples of your experiment to the *merged* arrays

Array name	Merged array
T-ALL-2-0h.CEL	1
T-ALL-2-8h.CEL	2
T-ALL-20-0h.CEL	1
T-ALL-20-8h.CEL	2

□ Average replicated genes (spots) within each microarray.

Averaging method

Choose the averaging method mean

□ log2 transform the data before averaging

Save merged values

Figure 2.15: Replicated microarrays / samples can be averaged (merged) in this page.

The replicate handling step for two color microarrays is shown in figure 2.9, in figure 2.15 the settings for the replicate handling of all other platforms are shown. The user can select the number of microarrays to which the microarrays of the experiments should be merged. In the example below the expression values of the Affymetrix GeneChip *T-ALL-2-0h.CEL* are averaged with those of *T-ALL-20-0h.CEL* and the same for the to chips with the patient samples after 8 hours GC treatment. As number of *merged arrays* 2 was selected, optional it is possible to give a (unique!) name to the merged arrays. Next the microarrays of the experiment have to be assigned to the merged arrays. Additionally it is possible to average the values of replicated spots (genes) within each microarray, or to log2 transform the data before averaging them and the method to calculate the average can be selected (mean or median value).

Chapter 3

Detection of differentially expressed genes

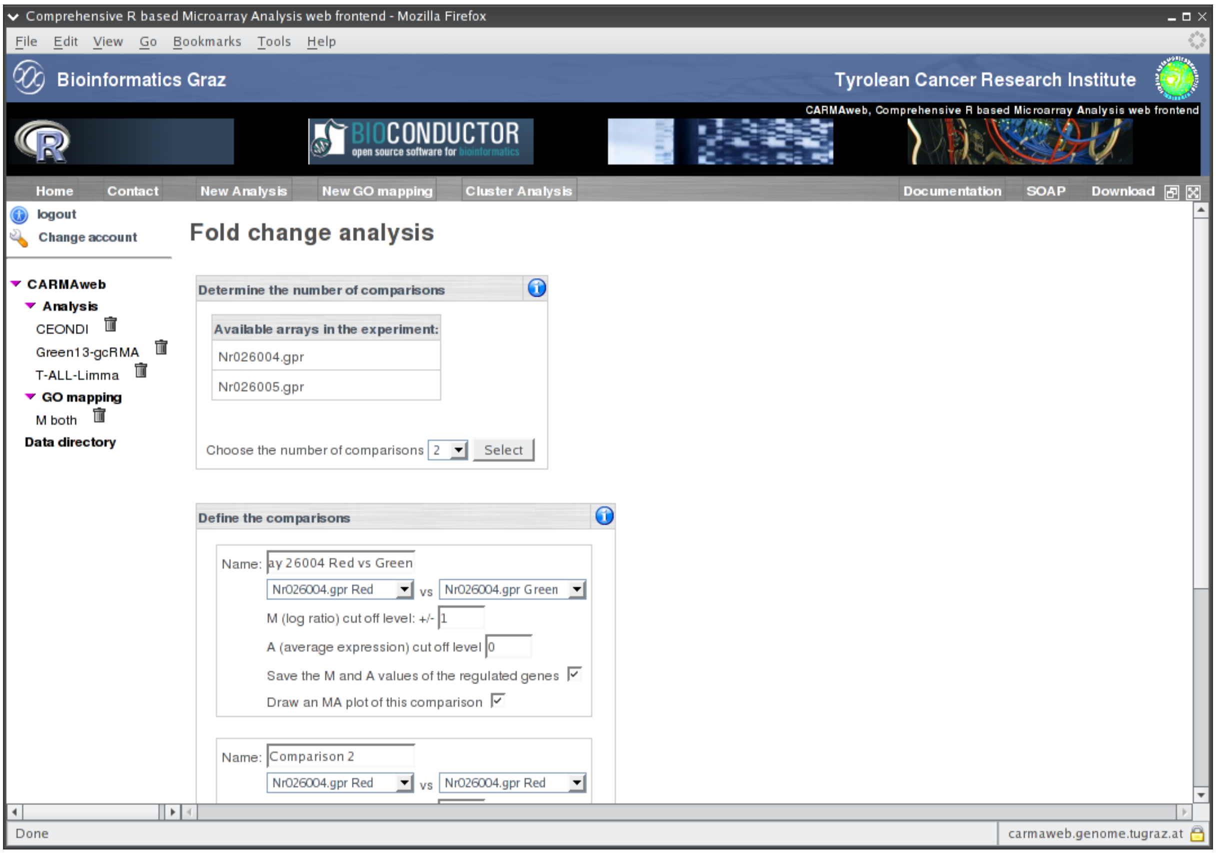
The two methods for the detection of differentially expressed genes described in the following sections can be applied to the normalized expression values of all three supported platforms (Affymetrix, two color microarrays, ABI microarrays) and also to any tabulator delimited text file containg numerical data..

3.1 Fold change analysis

For microarray experiments with a small number of arrays and thus samples, the detection for differentially expressed genes using a simple fold change analysis is possible. Differentially expressed genes are defined in this context as those genes that show a bigger increase or decrease in their expression, comparing the expression values of the sample to the control sample, than a user defined value. CARMAweb allows also to rank genes according to the number of comparisons in which they were defined to be differentially expressed.

CARMAweb allows to set cut–off values for both the M (regulation) and the A (average expression) value. In the example a M cut–of value of 1 was selected. Therefore all genes that show a two fold up– or down–regulation are defined to be differentially expressed.

The web page where comparisons can be defined is shown in figure 3.1. The number of comparisons for the microarray analysis has to be defined, then the sample (signal channel) and the control sample can be selected for each comparison. Genes that have a bigger expression value in the sample compared to the control sample lead to positive M values whereas negative M values derive from genes that are higher expressed in the control compared to the sample. Other parameters for a comparisons are the M and A value cut–off and the filename for the data table containing the regulated genes from the comparisons. Additionally a MA plot can be drawn for every comparison.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Fold change analysis

Determine the number of comparisons

Available arrays in the experiment:
Nr026004.gpr
Nr026005.gpr

Choose the number of comparisons 2 Select

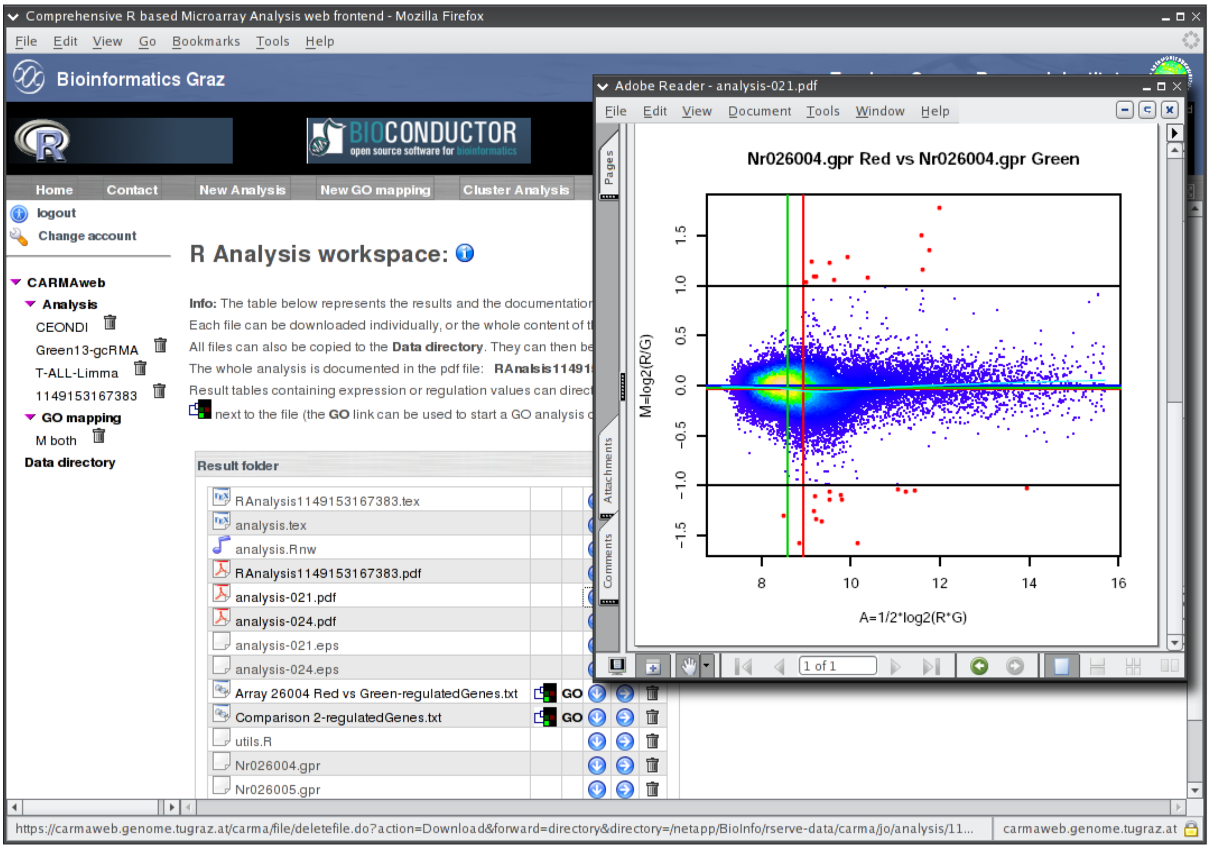
Define the comparisons

Name: ay 26004 Red vs Green
Nr026004.gpr Red vs Nr026004.gpr Green
M (log ratio) cut off level: +/- 1
A (average expression) cut off level 0
Save the M and A values of the regulated genes ☑
Draw an MA plot of this comparison ☑

Name: Comparison 2
Nr026004.gpr Red vs Nr026004.gpr Red

Figure 3.1: Foldchange analysis.

The result of such a fold change analysis is shown in figure 3.2. Again the different result files (analysis report file and tables) and figures can be displayed or downloaded individually, or the whole content of the result folder can be downloaded as a zip archive. The MA plot shown in figure 3.2 scatters the average expression values (A) against the differential expression values (M) of all genes between the two samples which are compared in the specific comparison. In this case the intensities measured in the red signal channel are compared against those from the green signal channel of a two-color microarray. Red points represent those genes that were defined to be differentially expressed between the two samples by the users cut-off criteria, which was in this case to define all genes that are more than two fold differentially expressed to be differentially expressed (M cut-off =1).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
BIOCONDUCTOR open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
1149153167383
GO mapping
M both
Data directory
Adobe Reader - analysis-021.pdf
File Edit View Document Tools Window Help
Pages
Attachments
Comments
Nr026004.gpr Red vs Nr026004.gpr Green
M=log2(R/G)
A=1/2*log2(R*G)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
8
10
12
14
16
1 of 1

R Analysis workspace:

Info: The table below represents the results and the documentation
Each file can be downloaded individually, or the whole content of t
All files can also be copied to the **Data directory**. They can then be
The whole analysis is documented in the pdf file: **RAnalsis11491**
Result tables containing expression or regulation values can direct
next to the file (the **GO** link can be used to start a GO analysis c

Result folder

RAnalysis1149153167383.tex	
analysis.tex	
analysis.Rnw	
RAnalysis1149153167383.pdf	
analysis-021.pdf	
analysis-024.pdf	
analysis-021.eps	
analysis-024.eps	
Array 26004 Red vs Green-regulatedGenes.txt	GO
Comparison 2-regulatedGenes.txt	GO
utils.R	
Nr026004.gpr	
Nr026005.gpr	

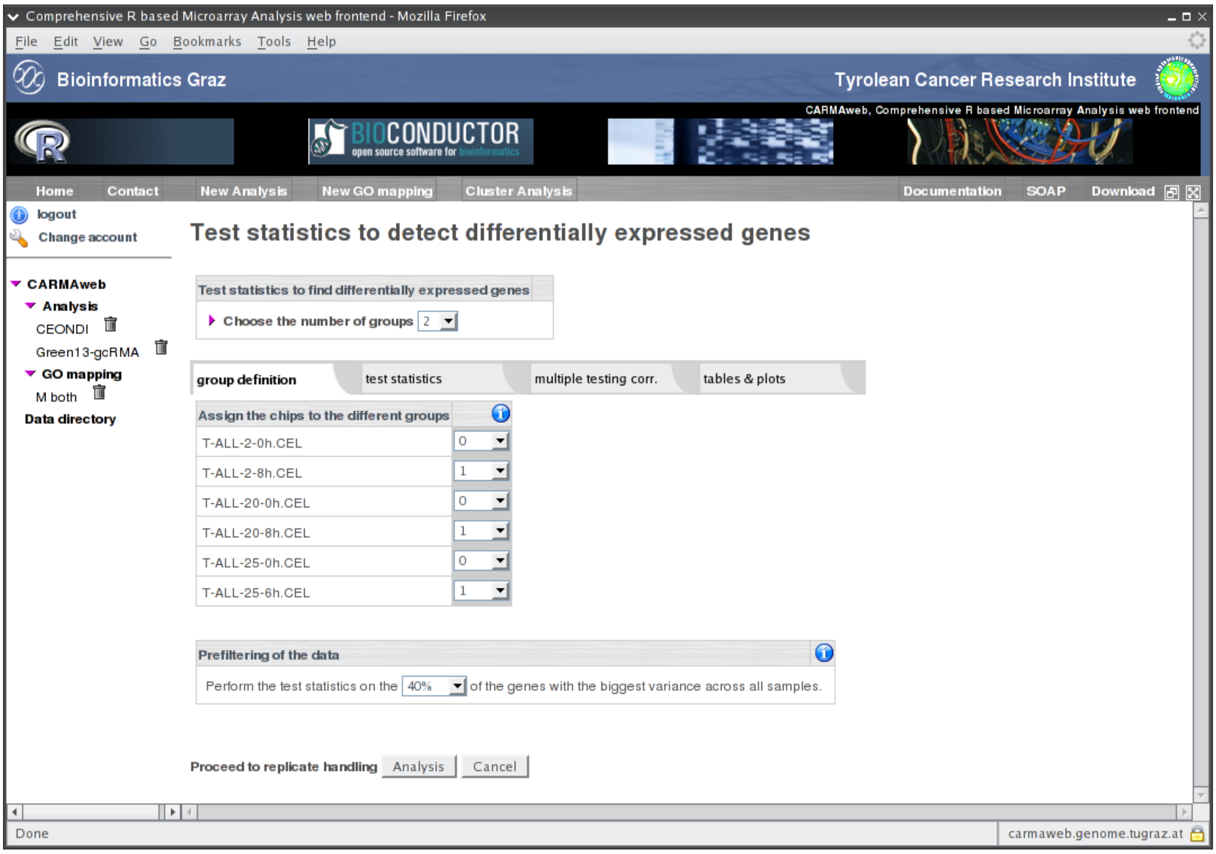
https://carmaweb.genome.tugraz.at/carma/file/deletefile.do?action=Download&forward=directory&directory=/netapp/BioInfo/rserve-data/carma/jo/analysis/11... carmaweb.genome.tugraz.at

Figure 3.2: Foldchange analysis result folder. The MA plot shown displays the average expression (A) against the differential expression (M) of all genes in two samples. Red points depict those genes that have been identified to be differentially expressed by the users cut-off criteria.

3.2 Test statistics

For microarray experiments with an appropriate number of arrays differentially expressed genes can be detected using statistical tests like the Mann Withney U test, the Students t test, the permutation (randomization) test, the *moderated t-statistics* from the limma package (based on a bayesian approach [13]) or SAM (Significance Analysis of Microarrays proposed by Tusher et al. [15]). The moderated t-statistics (empirical bayesian approach) are especially useful for small group sizes. For experiments where each sample has its own control sample (e.g. samples of patients before and after treatment) *paired* test statistics are suggested (either paired moderated t-test or paired permutation based test statistics). Microarray experiments generate large multiplicity problems in which thousands of hypothesis are tested simultaneously within one experiment [4] (e.g. is gene x differentially expressed between a treated and untreated sample). BioConductors multtest package provides suitable methods to adjust p values according to this multiple hypothesis testing problem. Available adjustment methods are for example the procedure introduced by Benjamini and Hochberg [2] for strong control of the FDR (false discovery rate, expected proportion of false positives among the rejected hypotheses) or the procedure by Westfall and Young [16] to control the FWER (family–wise error rate, probability of at least one false positive).

To alleviate the loss of power from the formidable multiplicity of gene-by-gene hypothesis testing within a microarray experiment, a non–specific prefiltering of the data can also be performed in CARMAweb (see figure 3.3 at the bottom of the page). Non–specific means without reference to the group the samples are into. The aim of this prefiltering step is to remove from consideration that set of genes that is not differentially expressed under any comparison. The web application allows the prefiltering of the data based on the variance within the data.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

Test statistics to detect differentially expressed genes

Test statistics to find differentially expressed genes

Choose the number of groups: 2

group definition | test statistics | multiple testing corr. | tables & plots

Assign the chips to the different groups	
T-ALL-2-0h.CEL	0
T-ALL-2-8h.CEL	1
T-ALL-20-0h.CEL	0
T-ALL-20-8h.CEL	1
T-ALL-25-0h.CEL	0
T-ALL-25-6h.CEL	1

Prefiltering of the data

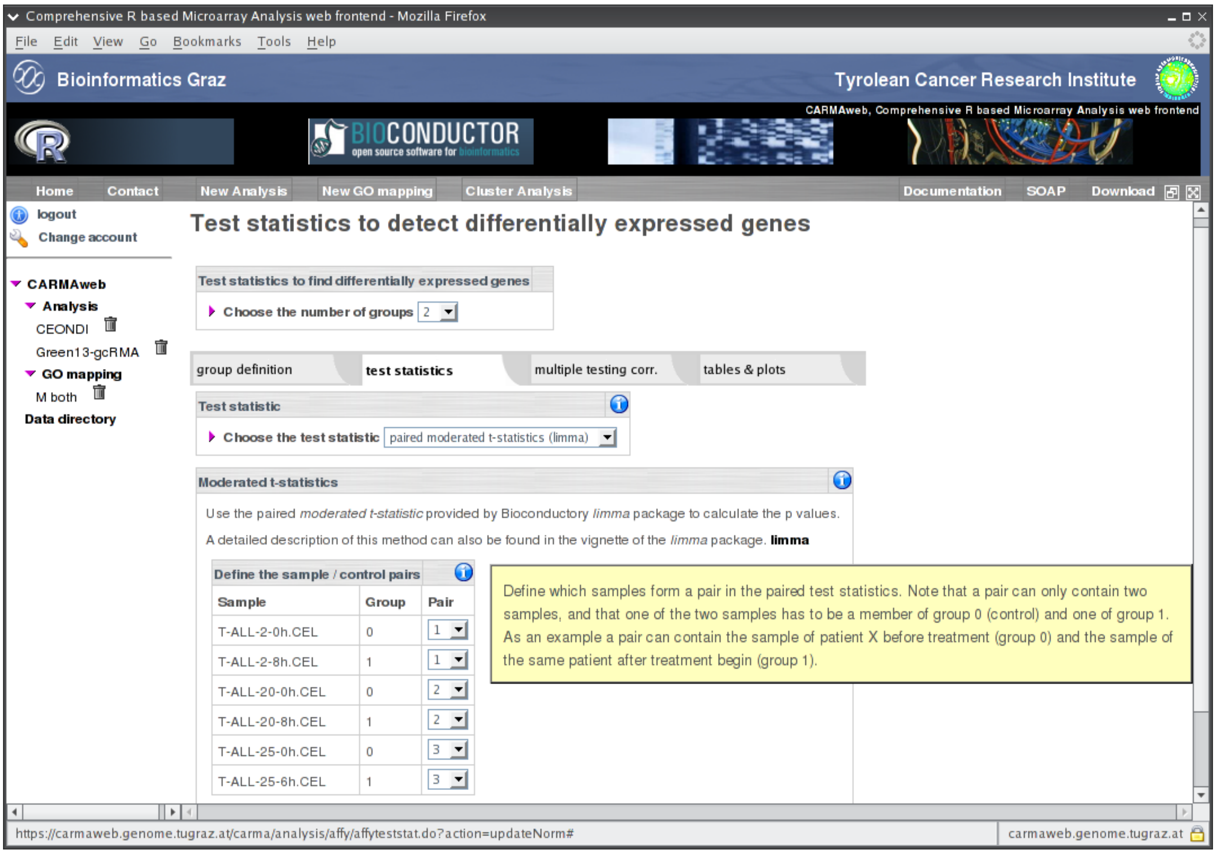
Perform the test statistics on the 40% of the genes with the biggest variance across all samples.

Proceed to replicate handling [Analysis] [Cancel]

Figure 3.3: Main page for the `multtest` package. Samples can be assigned to the different groups and the data set can be restricted to the x% of genes with the biggest variance.

The test analysis described in this section uses the normalized data from 6 Affymetrix GeneChips (the preprocessing of these chips is described in section 2.1). The samples are from three patients with childhood t-cell acute lymphoblastic leukemia (ALL), one sample before and one after treatment for 6 hours with dexametosone (glucocorticoids) respectively. To detect genes differentially expressed between the treated and untreated samples, the control samples have to be assigned to group 0 and the treated samples to group 1 (see figure 3.3). To alleviate the loss of power from the formidable multiplicity of gene-by-gene hypothesis testing a non-specific prefiltering is performed. For this purpose the further analysis is restricted to the 40% of the probesets with the biggest variance over all samples.

The test statistic can be selected on the page displayed in figure 3.4 (click on the *test statistics* tab to get there). By default a permutation based t test is selected. In our example we use *paired moderated t-test*, as we have paired data (sample from each patient before and after treatment), and the number of biological replicates is very small (the moderated t-test performs better on small group sizes compared to the other test statistics). For the paired moderated t-test the sample pairs have to be defined (each pair must have one sample from group 0 and oune from group 1).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
group definition
test statistics
multiple testing corr.
tables & plots
https://carmaweb.genome.tugraz.at/carma/analysis/affy/affyteststat.do?action=updateNorm#
carmaweb.genome.tugraz.at

Test statistics to detect differentially expressed genes

Test statistics to find differentially expressed genes

Choose the number of groups: 2

Test statistic

Choose the test statistic: paired moderated t-statistics (limma)

Moderated t-statistics

Use the paired *moderated t-statistic* provided by Bioconductory *limma* package to calculate the p values.

A detailed description of this method can also be found in the vignette of the *limma* package. **limma**

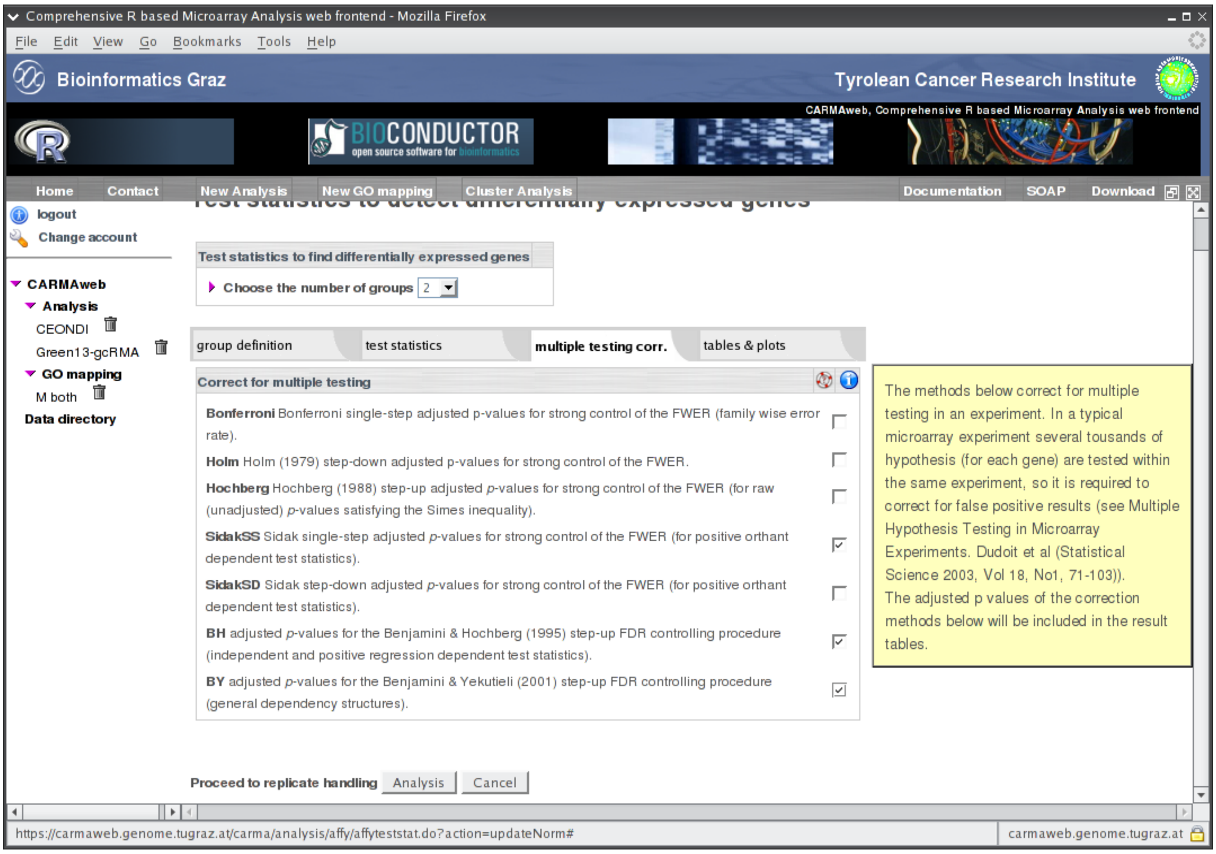
Define the sample / control pairs

Sample	Group	Pair
T-ALL-2-0h.CEL	0	1
T-ALL-2-8h.CEL	1	1
T-ALL-20-0h.CEL	0	2
T-ALL-20-8h.CEL	1	2
T-ALL-25-0h.CEL	0	3
T-ALL-25-6h.CEL	1	3

Define which samples form a pair in the paired test statistics. Note that a pair can only contain two samples, and that one of the two samples has to be a member of group 0 (control) and one of group 1. As an example a pair can contain the sample of patient X before treatment (group 0) and the sample of the same patient after treatment begin (group 1).

Figure 3.4: *test statistics* tab. The test statistic that should be used to detect the differentially expressed genes can be selected on this page. The pop-up window gives information about how sample pairs for paired test statistics can be defined. Such information pop–up windows are displayed when the user moves the mouse over an information button.

The method (or methods) that should be used to adjust for the multiple testing within the experiment can be selected on the page shown in figure 3.5 (tab *multiple testing corr.*).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
group definition
test statistics
multiple testing corr.
tables & plots
Proceed to replicate handling
Analysis
Cancel
https://carmaweb.genome.tugraz.at/carma/analysis/affy/affyteststat.do?action=updateNorm#
carmaweb.genome.tugraz.at

Test statistics to find differentially expressed genes

Choose the number of groups 2

Correct for multiple testing

Bonferroni Bonferroni single-step adjusted p-values for strong control of the FWER (family wise error rate). □

Holm Holm (1979) step-down adjusted p-values for strong control of the FWER. □

Hochberg Hochberg (1988) step-up adjusted *p*-values for strong control of the FWER (for raw (unadjusted) *p*-values satisfying the Simes inequality). □

SidakSS Sidak single-step adjusted *p*-values for strong control of the FWER (for positive orthant dependent test statistics). ☑

SidakSD Sidak step-down adjusted *p*-values for strong control of the FWER (for positive orthant dependent test statistics). □

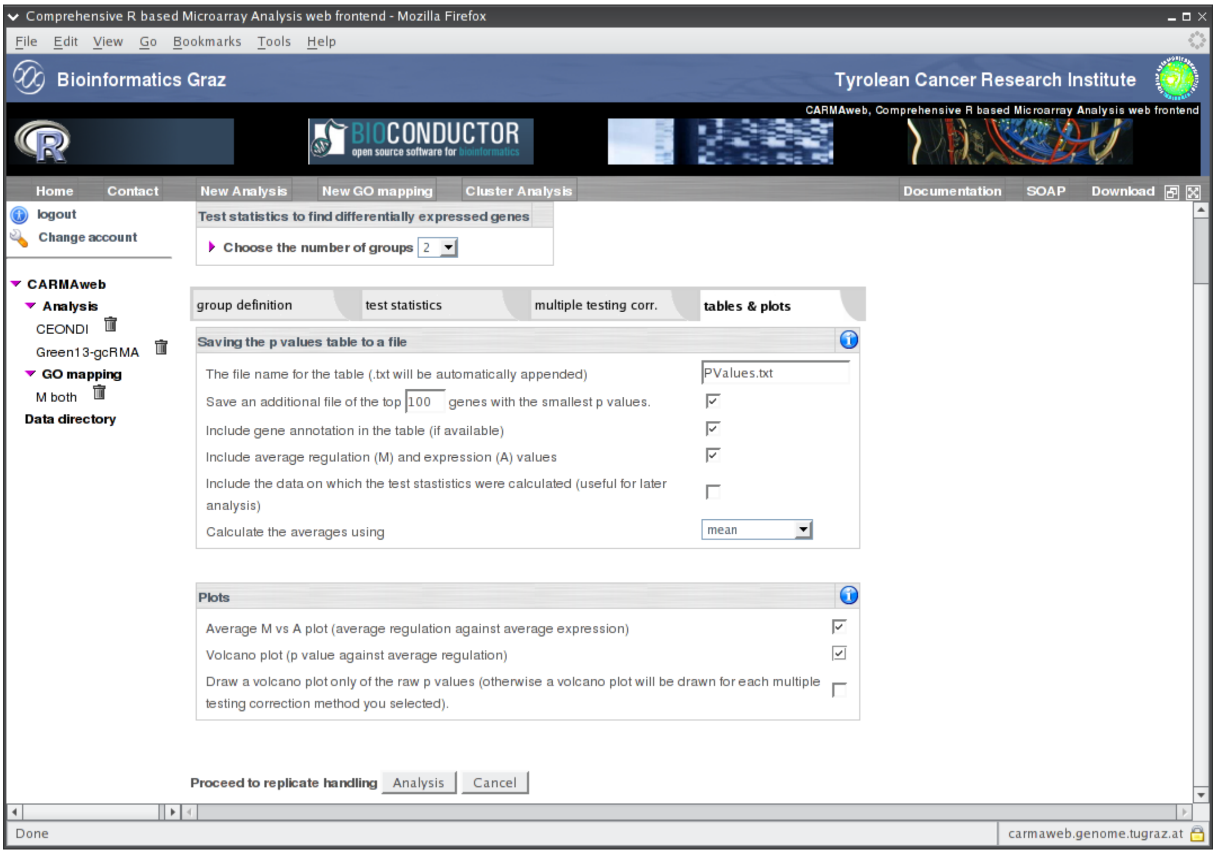
BH adjusted *p*-values for the Benjamini & Hochberg (1995) step-up FDR controlling procedure (independent and positive regression dependent test statistics). ☑

BY adjusted *p*-values for the Benjamini & Yekutieli (2001) step-up FDR controlling procedure (general dependency structures). ☑

The methods below correct for multiple testing in an experiment. In a typical microarray experiment several tousands of hypothesis (for each gene) are tested within the same experiment, so it is required to correct for false positive results (see Multiple Hypothesis Testing in Microarray Experiments. Dudoit et al (Statistical Science 2003, Vol 18, No1, 71-103)). The adjusted p values of the correction methods below will be included in the result tables.

Figure 3.5: *multiple testing correction* tab. The method (or methods) that should be used to adjust for the multiple testing problem can be selected on this page.

Finally the plots and tables that should be created can be defined on the 4th tab (*tables & plots*) shown in figure 3.6. For each analysis a tab delimited text file containing the p values calculated for each gene in the data set is created by default.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
Tyrolean Cancer Research Institute
CARMAweb, Comprehensive R based Microarray Analysis web frontend
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
GO mapping
M both
Data directory
group definition
test statistics
multiple testing corr.
tables & plots
Done
carmaweb.genome.tugraz.at

Test statistics to find differentially expressed genes

Choose the number of groups 2

Saving the p values table to a file

The file name for the table (.txt will be automatically appended) PValues.txt

Save an additional file of the top 100 genes with the smallest p values. ☑

Include gene annotation in the table (if available) ☑

Include average regulation (M) and expression (A) values ☑

Include the data on which the test stastistics were calculated (useful for later analysis) □

Calculate the averages using mean

Plots

Average M vs A plot (average regulation against average expression) ☑

Volcano plot (p value against average regulation) ☑

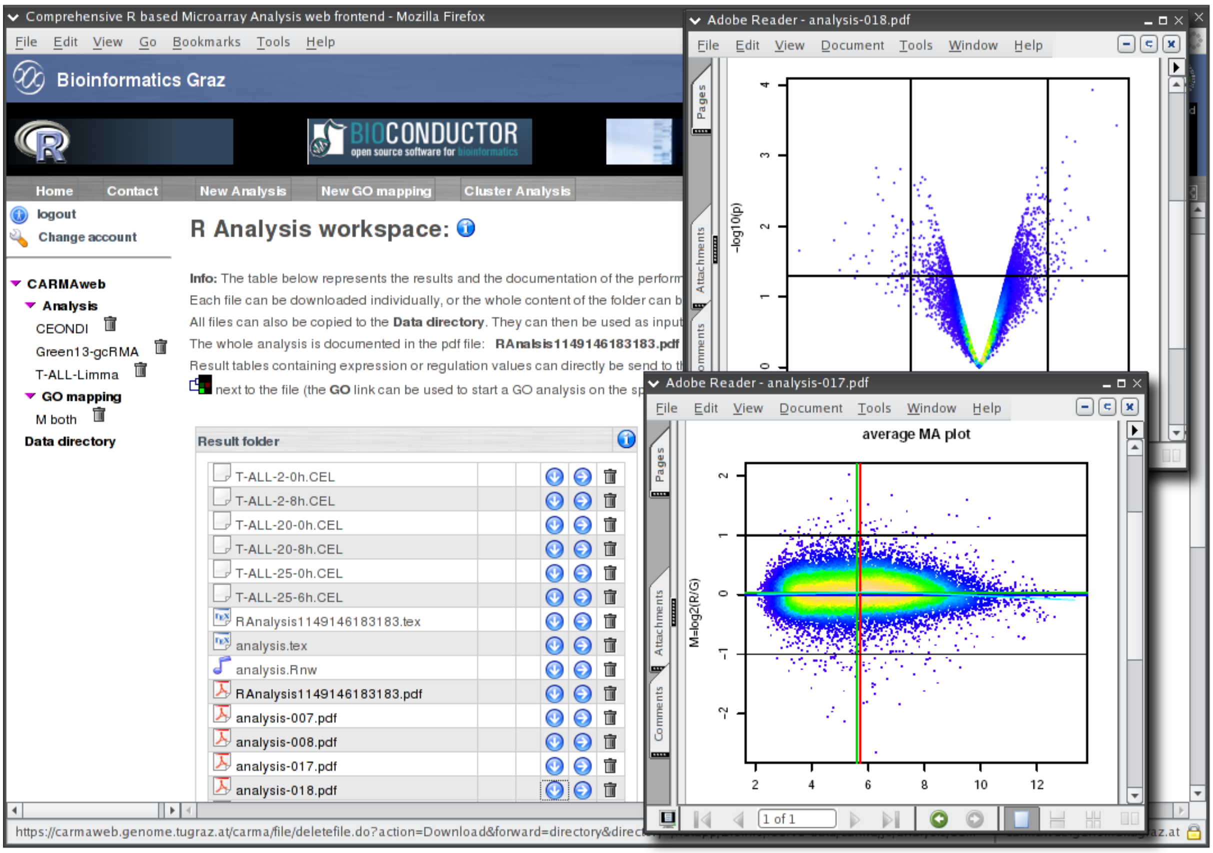
Draw a volcano plot only of the raw p values (otherwise a volcano plot will be drawn for each multiple testing correction method you selected). □

Proceed to replicate handling Analysis Cancel

Figure 3.6: *plots & tables* tab. The plots and tables that should be created from the data can be defined using this page.

The results of the test analysis performed in this chapter are shown in figure 3.7. The MA plot that is shown scatters the average expression of all probesets between the two groups on the x-axis (A values) against the differential expression on the y-axis (M values). In fact, as this is a average MA plot, it scatters the mean M versus the mean A values, where the mean A and M values are calculated from the mean expression values of each probeset in group 1 against the mean expression values in group 0.

The volcano plot scatters the mean M values (x-axis) against the test statistic (more in detail against the -log10 of the p-values). Points (probesets) with small p-values are in the upper region of the plot (have large values on the y axis), whereas probesets that are not significantly differentially expressed are near y=0.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
BIOCONDUCTOR open source software for bioinformatics
Home Contact New Analysis New GO mapping Cluster Analysis
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Adobe Reader - analysis-018.pdf
File Edit View Document Tools Window Help
Pages
Attachments
Comments
-log10(p)
0
1
2
3
4
Adobe Reader - analysis-017.pdf
File Edit View Document Tools Window Help
average MA plot
M=log2(R/G)
-2
-1
0
1
2
4
6
8
10
12
1 of 1
https://carmaweb.genome.tugraz.at/carma/file/deletefile.do?action=Download&forward=directory&direc
z.at

R Analysis workspace:

Info: The table below represents the results and the documentation of the perform
Each file can be downloaded individually, or the whole content of the folder can b
All files can also be copied to the **Data directory**. They can then be used as input
The whole analysis is documented in the pdf file: **RAnalsis1149146183183.pdf**
Result tables containing expression or regulation values can directly be send to th
next to the file (the **GO** link can be used to start a GO analysis on the sp

Result folder

T-ALL-2-0h.CEL
T-ALL-2-8h.CEL
T-ALL-20-0h.CEL
T-ALL-20-8h.CEL
T-ALL-25-0h.CEL
T-ALL-25-6h.CEL
RAnalysis1149146183183.tex
analysis.tex
analysis.Rnw
RAnalysis1149146183183.pdf
analysis-007.pdf
analysis-008.pdf
analysis-017.pdf
analysis-018.pdf

Figure 3.7: Result page for a detection of differentially expressed genes analysis using test statistics. Like in all result pages the different figures and tables can be opened or downloaded individually, or the whole content of the result folder can be downloaded as a zip archive. The two figures shown are a average MA plot (group 1 against group 0) and a volcano plot of the raw p-values.

CARMAweb allows to add more *detection–of–differentially–expressed–genes–steps* to one microarray analsis (after adding one step to the analysis on the analysis page (see figure 2.3 in section 2.1)).

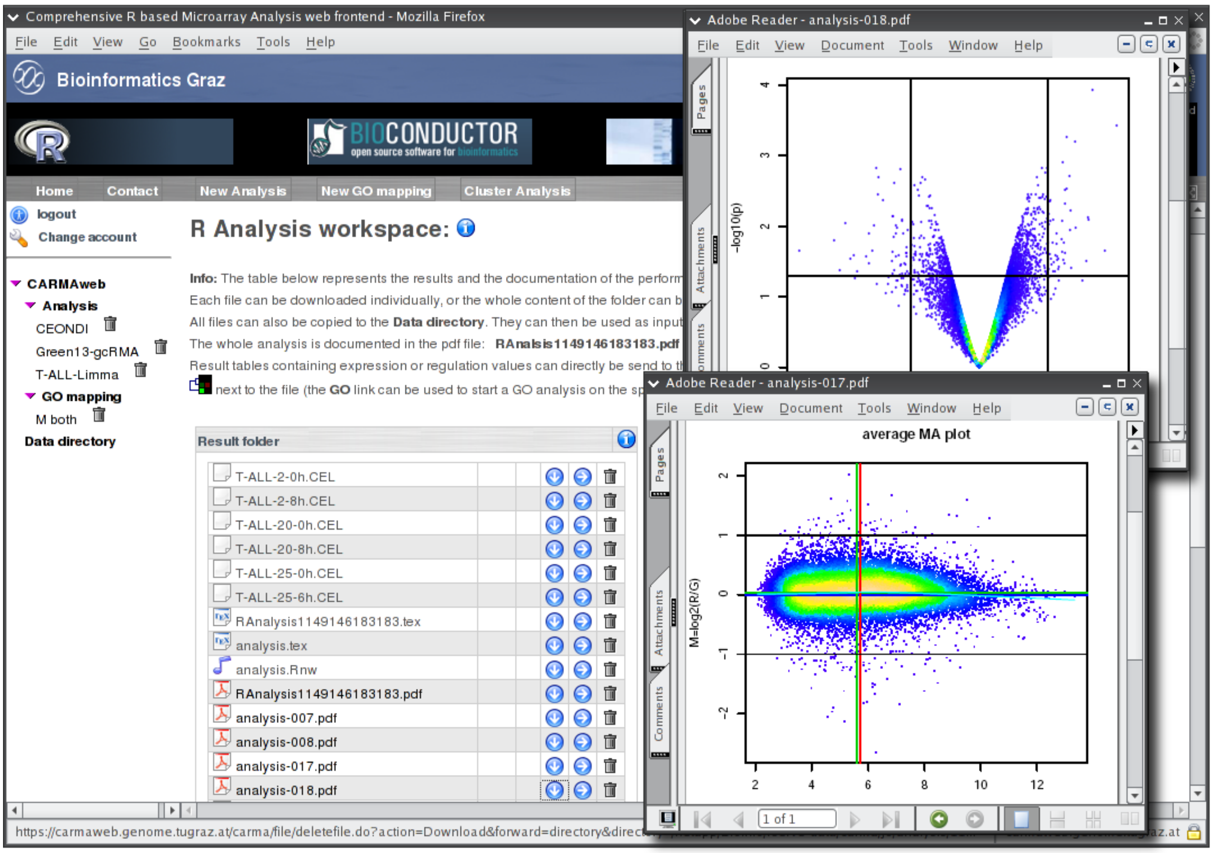
Chapter 4

Analysis result folder

Once the calculation of an analysis is finished its results are accessible through the *R Analysis workspace* page (figure 4.1). The content of this folder can be downloaded as a zip archive, or each single file can be downloaded separately (by clicking on the ⬇ icon next to the file). Additionally all files from the result folder (*R Analysis workspace*) can be copied to the users *Data directory* using the ➔ icon or be deleted by clicking on the 🗑 icon or directly sent to the cluster analysis module GenesisWeb using the ▦ icon. Copying files from an analysis result folder to the users *Data directory* allow to use this result files as input files for further anlyses (e.g. when a file with the normalized expression values is copied to the *Data directory* this can be used for a new analsis using normalized data (=detection of differentially expressed genes). Result files can then also be used as input files for a cluster analysis or a GO analysis).

For each analysis a report file is generated, in which all performed analysis steps and the descriptions for the various methods used in the analysis are delineated, thus guarantee a maximal transparency and reproducibility of the analysis and allowing the user to re–perform the whole analysis on a local workstation by typing the R commands from the analysis report into a R console. Depending on the users choices different tables with expression or regulation values are created. This may range from tables containing the normalized expression values of all genes in all arrays, to tables with regulation or expression values of the subsets of differentially expressed genes or tables containing the raw p values and p values adjusted using the various adjustment methods.

All plots are embedded in the analysis report file and are additionally available as single pdf (end eps) files.



Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Bioinformatics Graz
BIOCONDUCTOR
open source software for bioinformatics
Home
Contact
New Analysis
New GO mapping
Cluster Analysis
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Adobe Reader - analysis-018.pdf
File Edit View Document Tools Window Help
Pages
Attachments
Comments
-log10(p)
0
1
2
3
4
Adobe Reader - analysis-017.pdf
File Edit View Document Tools Window Help
average MA plot
M=log2(R/G)
-2
-1
0
1
2
4
6
8
10
12
1 of 1

R Analysis workspace:

Info: The table below represents the results and the documentation of the perform
Each file can be downloaded individually, or the whole content of the folder can b
All files can also be copied to the **Data directory**. They can then be used as input
The whole analysis is documented in the pdf file: **RAnalsis1149146183183.pdf**
Result tables containing expression or regulation values can directly be send to t
next to the file (the **GO** link can be used to start a GO analysis on the sp

Result folder

T-ALL-2-0h.CEL
T-ALL-2-8h.CEL
T-ALL-20-0h.CEL
T-ALL-20-8h.CEL
T-ALL-25-0h.CEL
T-ALL-25-6h.CEL
RAnalysis1149146183183.tex
analysis.tex
analysis.Rnw
RAnalysis1149146183183.pdf
analysis-007.pdf
analysis-008.pdf
analysis-017.pdf
analysis-018.pdf

https://carmaweb.genome.tugraz.at/carma/file/deletefile.do?action=Download&forward=directory&direc

Figure 4.1: Result page for a detection of differentially expressed genes analysis using test statistics. Like in all result pages the different figures and tables can be opened or downloaded individually, or the whole content of the result folder can be downloaded as a zip archive. The two figures shown are a average MA plot (group 1 against group 0) and a volcano plot of the raw p-values. All points in the figures are colored according to the local points density (yellow represents high density, blue low).

Chapter 5

GO analysis

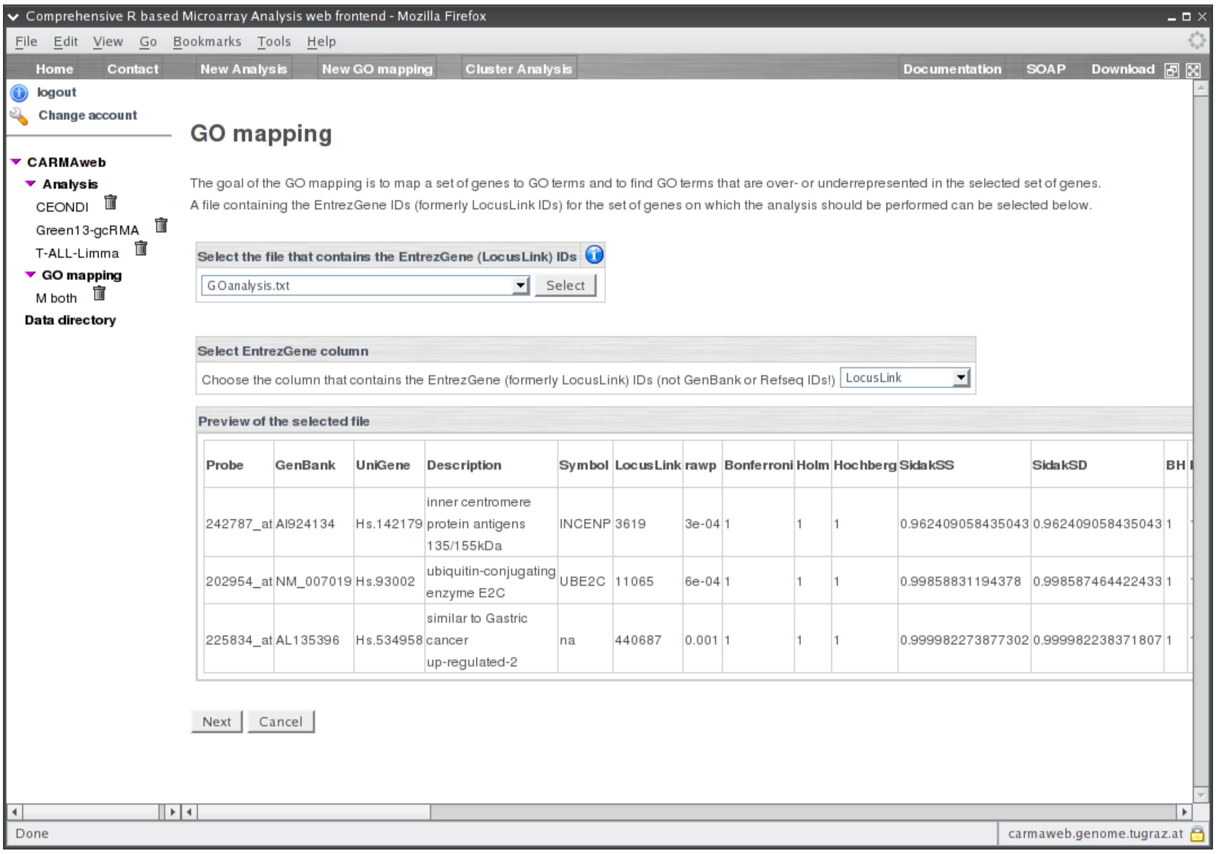
A Gene Ontology is a structured vocabulary to describe gene and gene products. In other words a gene ontology is a common set of words and descriptions that are related to each other. The relationships of the words (GO terms) of the ontology are all many to many. This means that each GO term can have a set of child terms and can also have a set of parent terms, where always child terms are more specific than parent terms. The Gene Ontolgy Consortium has build three independent ontologies to describe the cellular component, molecular function or biological process of gene products.

The goal of GO analysis is to assist in the biological interpretation of microarray analyses by providing information about the biological process, molecular function or cellular component a set of genes of interest is involved in. The algorithm of the GO analysis firts creates the induced GO graph for the set of interesting genes (provided by the input file). This GO graph contains all GO terms one or more genes of interest can be associated with. Next it defines for each GO term the number of genes of interest that can be mapped to the specific GO term, as well as the number of genes that can be detected with the microarray used in the experiment and that can be associated with the GO term. Based on this data p-values can then be calculated that provide the information if the genes of interest are over-represented in the specific GO term.

The GO analysis in `CARMAweb` is performed by `BioConductors` `GOstats` and `GO` packages.

As input file any tab delimited file can be used, that contains in one column the Entrez Gene (formerly LocusLink) identifiers of the genes of interest. A new GO analysis can be started using the link *New GO mapping* in the button bar of the main page. The user has then to select one of the files that are in his CARMAweb *Data directory* for the analysis (keep in mind that the files have to contain one column with the Entrez Gene IDs of the genes of interest, the probe sets of the Affymetrix GeneChip analysis in CARMAweb are usually automatically annotated to Entrez Gene identifiers, so result files from these analyses contain already one column with Entrez Gene IDs).

The following GO mapping analysis is performed using the test file called *GOanalysis.txt*. First the file containing the corresponding Entrez Gene IDs has to be selected (in our example the *GOanalysis.txt* file). After selecting the correct input file the first four lines of this file are read and displayed on the page. This should ease the correct selection of the column containing the Entrez Gene IDs. Then the column containing the Entrez Gene (LocusLink) identifiers has to be selected which is in our example called *LocusLink* (see figure 5.1).


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Select
Next
Cancel
Done
carmaweb.genome.tugraz.at

GO mapping

The goal of the GO mapping is to map a set of genes to GO terms and to find GO terms that are over- or underrepresented in the selected set of genes.
A file containing the EntrezGene IDs (formerly LocusLink IDs) for the set of genes on which the analysis should be performed can be selected below.

Select the file that contains the EntrezGene (LocusLink) IDs

GOanalysis.txt

Select EntrezGene column

Choose the column that contains the EntrezGene (formerly LocusLink) IDs (not GenBank or Refseq IDs!) LocusLink

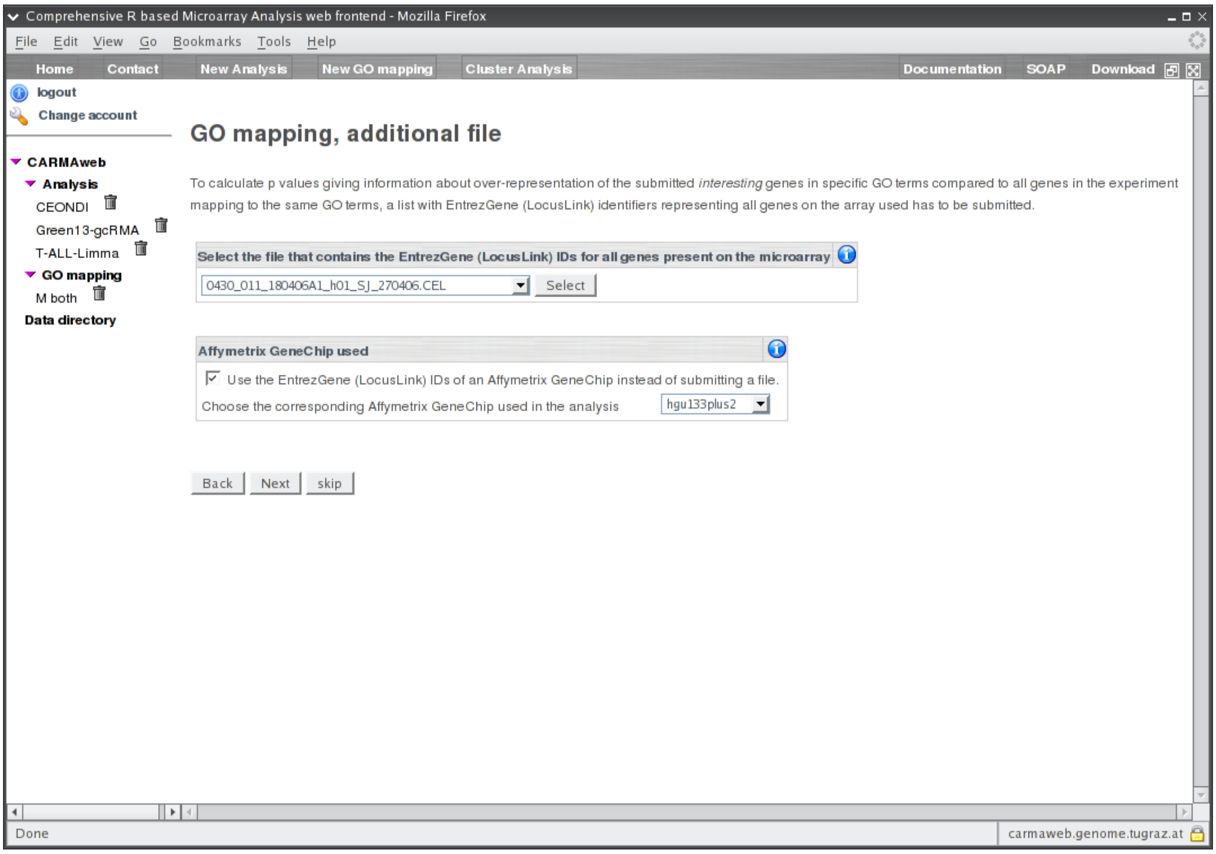
Preview of the selected file

Probe	GenBank	UniGene	Description	Symbol	LocusLink	rawp	Bonferroni	Holm	Hochberg	SidakSS	SidakSD	BH
242787_at	AI924134	Hs.142179	inner centromere protein antigens 135/155kDa	INCENP	3619	3e-04	1	1	1	0.962409058435043	0.962409058435043	1
202954_at	NM_007019	Hs.93002	ubiquitin-conjugating enzyme E2C	UBE2C	11065	6e-04	1	1	1	0.99858831194378	0.998587464422433	1
225834_at	AL135396	Hs.534958	similar to Gastric cancer up-regulated-2	na	440687	0.001	1	1	1	0.999982273877302	0.999982238371807	1

Figure 5.1: GO analysis. Selection of the file containing the Entrez Gene identifiers of the genes of interest, and the column in the selected file containing those IDs.

To calculate hypergeometric p–values that give information about the over–representation of the genes of interest in the GO terms compared to all genes that can be mapped to the corresponding terms, a file containing all Entrez Gene identifier of all genes that can be detected with the microarray used has to be selected (figure 5.2). Affymetrix users can choose the Affymetrix GeneChip used in the analysis and do not have to upload or select a file. By skipping this step no (correct) p values will be calculated and only the GO graph is generated.
The genes of the test file were detected using Affymetrix GeneChip *hgu133plus2*, therefore this chip has been selected (figure 5.2).

The ontologies for which the GO analysis should be performed can be selected on the page shown in figure 5.3. Once the desired parameters and settings have been chosen the analysis can be started by clicking on the *Start* button.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Done
carmaweb.genome.tugraz.at

GO mapping, additional file

To calculate p values giving information about over-representation of the submitted *interesting* genes in specific GO terms compared to all genes in the experiment mapping to the same GO terms, a list with EntrezGene (LocusLink) identifiers representing all genes on the array used has to be submitted.

Select the file that contains the EntrezGene (LocusLink) IDs for all genes present on the microarray

0430_011_180406A1_h01_SJ_270406.CEL [Select]

Affymetrix GeneChip used

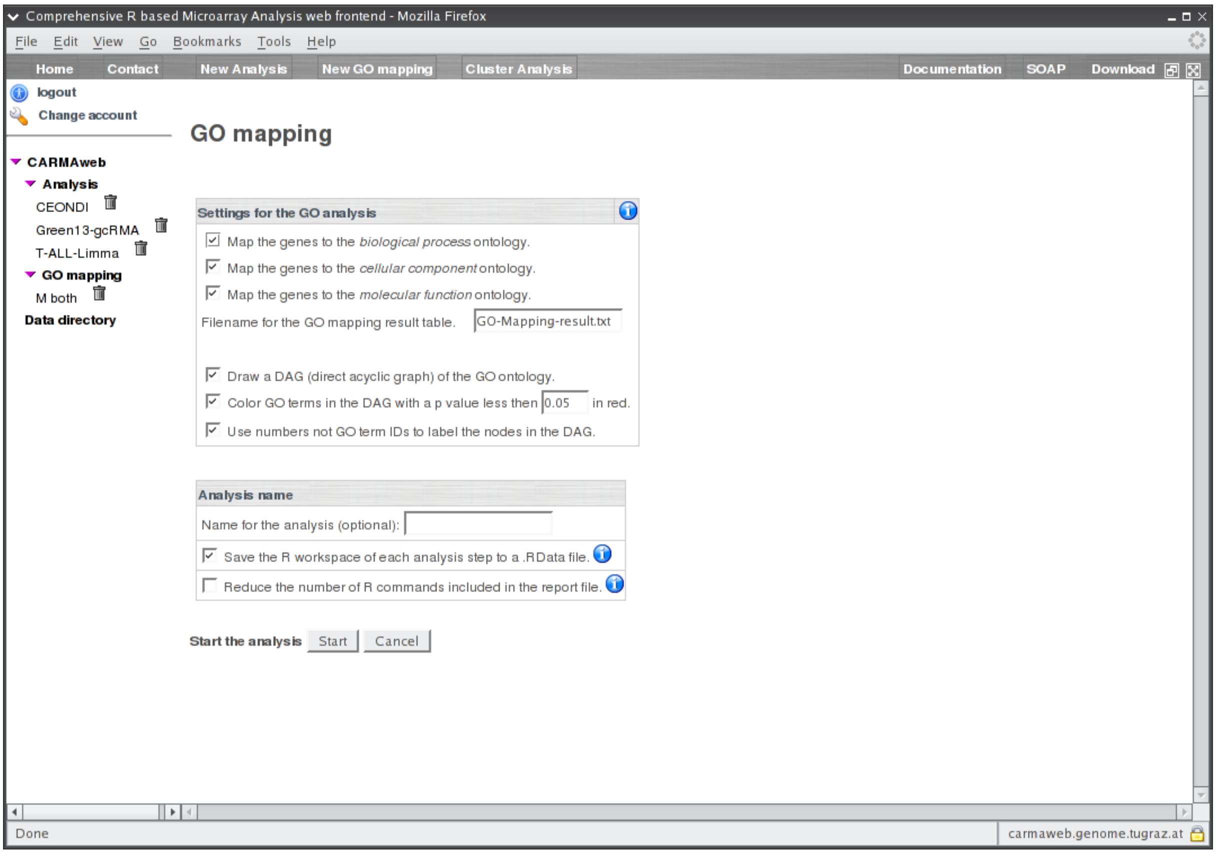
☑ Use the EntrezGene (LocusLink) IDs of an Affymetrix GeneChip instead of submitting a file.

Choose the corresponding Affymetrix GeneChip used in the analysis: hgu133plus2

[Back] [Next] [skip]

Figure 5.2: GO analysis. Selection of the file that contains the Entrez Gene identifiers of all genes on the microarray used. Affymetrix users can select the GeneChip that was used in the microarray experiment.

The analysis results are available through the results page (figure 5.4). For each ontology a GO graph and a table containing all the GO terms, to which the genes of interest can be mapped to, are created. The table contains the GO terms together with the number of genes of interest that can be mapped to the GO term, the number of genes on the array that can be mapped to the GO term and the total number of genes that correspond to the GO term (from every species). The row numbers of the GO terms in this table are used as the labels of the GO terms in the corresponding GO graph.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
Data directory
Start the analysis
Start
Cancel
Done
carmaweb.genome.tugraz.at

GO mapping

Settings for the GO analysis

☑ Map the genes to the *biological process* ontology.

☑ Map the genes to the *cellular component* ontology.

☑ Map the genes to the *molecular function* ontology.

Filename for the GO mapping result table. GO-Mapping-result.txt

☑ Draw a DAG (direct acyclic graph) of the GO ontology.

☑ Color GO terms in the DAG with a p value less then 0.05 in red.

☑ Use numbers not GO term IDs to label the nodes in the DAG.

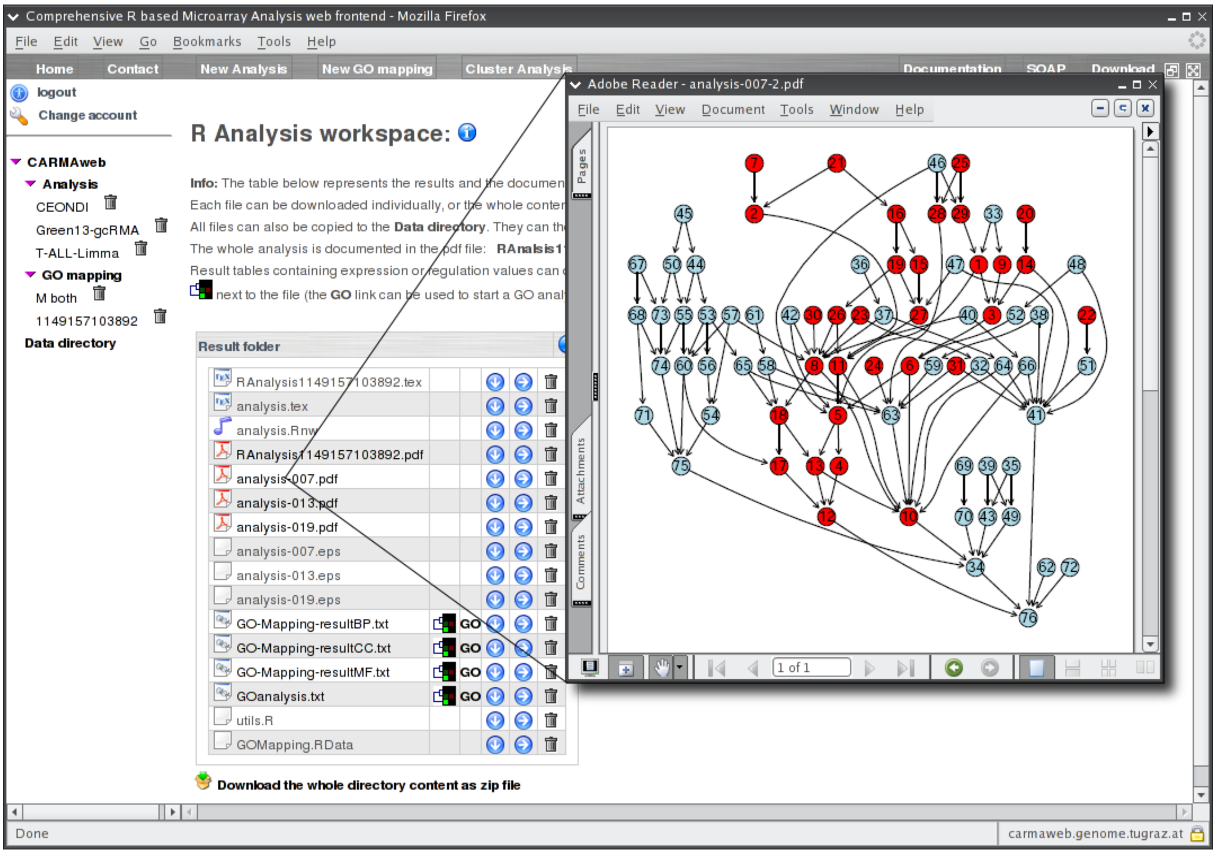
Analysis name

Name for the analysis (optional):

☑ Save the R workspace of each analysis step to a .RData file.

□ Reduce the number of R commands included in the report file.

Figure 5.3: GO analysis. Additional parameters for the GO analysis.


Comprehensive R based Microarray Analysis web frontend - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Home Contact New Analysis New GO mapping Cluster Analysis Documentation SOAP Download
logout
Change account
CARMAweb
Analysis
CEONDI
Green13-gcRMA
T-ALL-Limma
GO mapping
M both
1149157103892
Data directory
Adobe Reader - analysis-007-2.pdf
File Edit View Document Tools Window Help
Pages
Attachments
Comments
1 of 1
Done
carmaweb.genome.tugraz.at

R Analysis workspace:

Info: The table below represents the results and the documen

Each file can be downloaded individually, or the whole conter

All files can also be copied to the **Data directory**. They can th

The whole analysis is documented in the pdf file: **RAnalsis1**

Result tables containing expression or regulation values can

next to the file (the **GO** link can be used to start a GO anal

Result folder

RAnalysis1149157103892.tex		
analysis.tex		
analysis.Rnw		
RAnalysis1149157103892.pdf		
analysis-007.pdf		
analysis-013.pdf		
analysis-019.pdf		
analysis-007.eps		
analysis-013.eps		
analysis-019.eps		
GO-Mapping-resultBP.txt		GO
GO-Mapping-resultCC.txt		GO
GO-Mapping-resultMF.txt		GO
GOanalysis.txt		GO
utils.R		
GOMapping.RData		

Download the whole directory content as zip file

Figure 5.4: GO analysis. GO analysis result folder. The GO graph for the cellular component ontology is shown in the additional window.

Chapter 6

Cluster analysis

The cluster analysis can be performed using CARMAwebs cluster analysis module `GenesisWeb` which can be used as a standalone web application (accessible at https://carmaweb.genome.tugraz.at/genesis) or in combination with CARMAweb. GenesisWeb uses facilities of the GenesisServer application server [14] to perform the cluster analysis calculations.

The layout of the GenesisWeb main page can be splitted up into the three parts button bar, navigation tree and center frame (see figure 6.2). The button bar can be used to manipulate the opened data set (for example using the *Normalize Data* menu, that allows to normalize the data (for example mean centering of the data)), or by starting a new cluster analysis by selecting one of the cluster analysis methods available. Cluster analysis algorithms implemented in GenesisWeb include the unsupervised methods HCL (hierarchical clustering), SOM (self organizing maps), KMC (k–means clustering). Additionally dimension reduction and visualization methods like the PCA (principle component analysis) and CA (correspondence analysis) can be applied to the data. For the classification of microarray data the supervised method SVM (support vector machines) is available. Algorithm specific parameters as well as the distance (similarity) measurement method can be selected by the user.

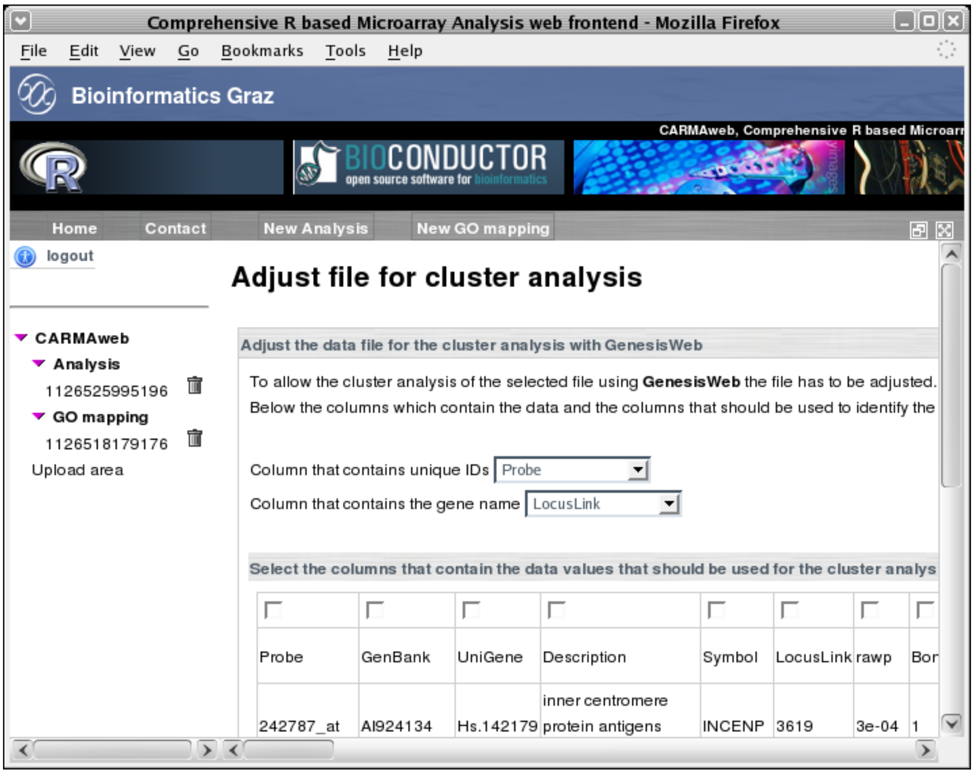
The navigation tree in the left part of the web page allows to open or upload new data files and to access cluster analysis results.

GenesisWeb takes as input files any tabulator delimited table can be used. The columns should correspond to the samples and the rows to the genes.

A new cluster analysis on files in the users *Data directory* can be started using the `New Cluster Analysis` button in the button bar of CARMAweb. A cluster analysis using GenesisWeb can also be started from CARMAweb directly by clicking on the ▦ icon next to a file (tab delimited txt file) in a analysis result page (see figure 4.1).

Cluster results (clusters exported to txt files) can be fetched back to the users *Data directory* using the corresponding form in the *Data directory* and thus be used for any further analysis (GO analysis...). To be able to perform a GO analysis on cluster analysis results it is required, that the data file used for the cluster analysis contains the Entrez Gene (LocusLink) IDs of the genes (either in the *Id* or *Name* column).

In the example below the input file for the GO analysis used in chapter 5 was used. This file contains also the normalized expression values of the 100 genes over a big number of samples (the columns containing the normalized expression values have column names beginning with *B–ALL*). By clicking on the file with the name *GOanalysis.txt* shown in the result folder in figure 5.4, a page is opened (figure 6.1) where the user can select the columns of the data file that contain the data that should be clustered (in the example those columns with column names beginning with *B–ALL*). Additionally the user should select the column that contains a unique identifier for each gene (in the example the *ProbeID* column). Based on the users selections the data file is generated and opened directly in GenesisWeb (see figure 6.2) by clicking on the next button at the bottom of the page.



Adjust file for cluster analysis

Adjust the data file for the cluster analysis with GenesisWeb

To allow the cluster analysis of the selected file using **GenesisWeb** the file has to be adjusted. Below the columns which contain the data and the columns that should be used to identify the

Column that contains unique IDs: Probe

Column that contains the gene name: LocusLink

Select the columns that contain the data values that should be used for the cluster analys

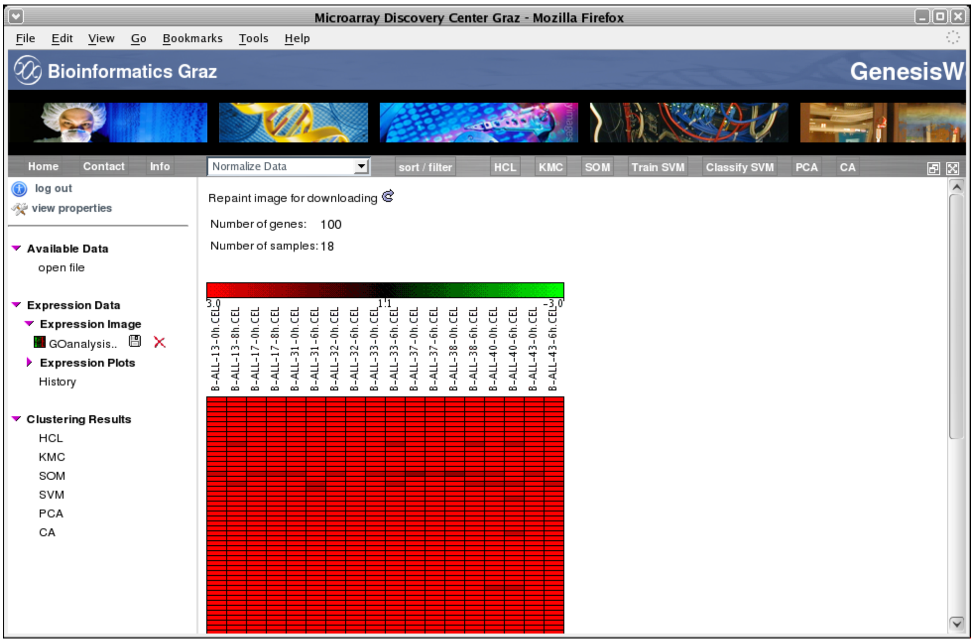
□	□	□	□	□	□	□	□
Probe	GenBank	UniGene	Description	Symbol	LocusLink	rawp	Bor
242787_at	AI924134	Hs.142179	inner centromere protein antigens	INCENP	3619	3e-04	1

Figure 6.1: Cluster analysis. The columns containing the identifiers for each gene, as well as the columns that contain the data that should be used for the cluster analysis have to be defined.

In the example this dataset was *mean centered* using the *Normalize Data* menu in the button bar of GenesisWeb and a hierarchical cluster analysis (euclidean distance measurement and clustering over the samples and the genes) is started by clicking on the HCL link in the button bar.

The result of a cluster analysis result using the HCL algorithm is shown in figure 6.3.

Gene and / or sample clusters can be selected and colored as well as exported to text files (these files contain



Repaint image for downloading

Number of genes: 100

Number of samples: 18

Figure 6.2: Cluster analysis. The cluster analysis module GenesisWeb. The button bar in the top of the page can be used to apply the different cluster analysis algorithms to the data. Analysis results as well as data files are accessible through the navigation tree in the left part of the web page.

the expression values of the selected genes). By clicking on the appropriate node in the hierarchical tree in the result shown in figure 6.3 a web page is opened (shown in figure 6.4) where the cluster color can be selected or the cluster can be exported to a text file.

By making a cluster *public* (using the web page shown in figure 6.4) all the genes (or samples) of the corresponding cluster are also labeled with the same color in other cluster results (for example using a different clustering algorithm) of the same data set. This is shown in figure 6.6, where the genes from the green marked cluster (see figure 6.5) are also labeled green in a PCA clustering result.

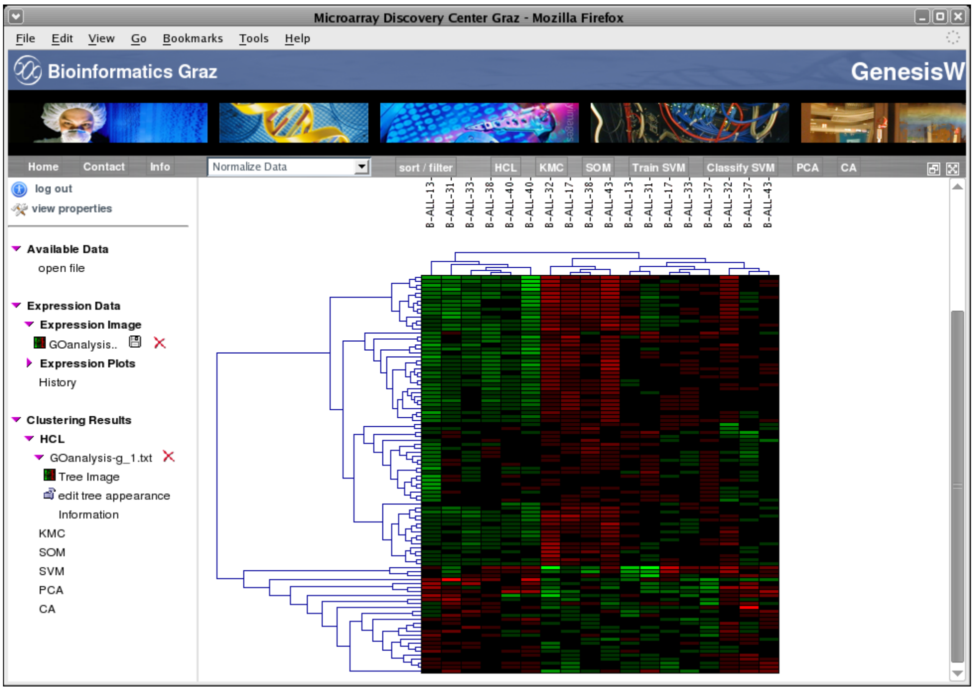


Figure 6.3: Result of a HCL cluster analysis of the test file. Gene and Sample clusters can be selected, colored and exported.

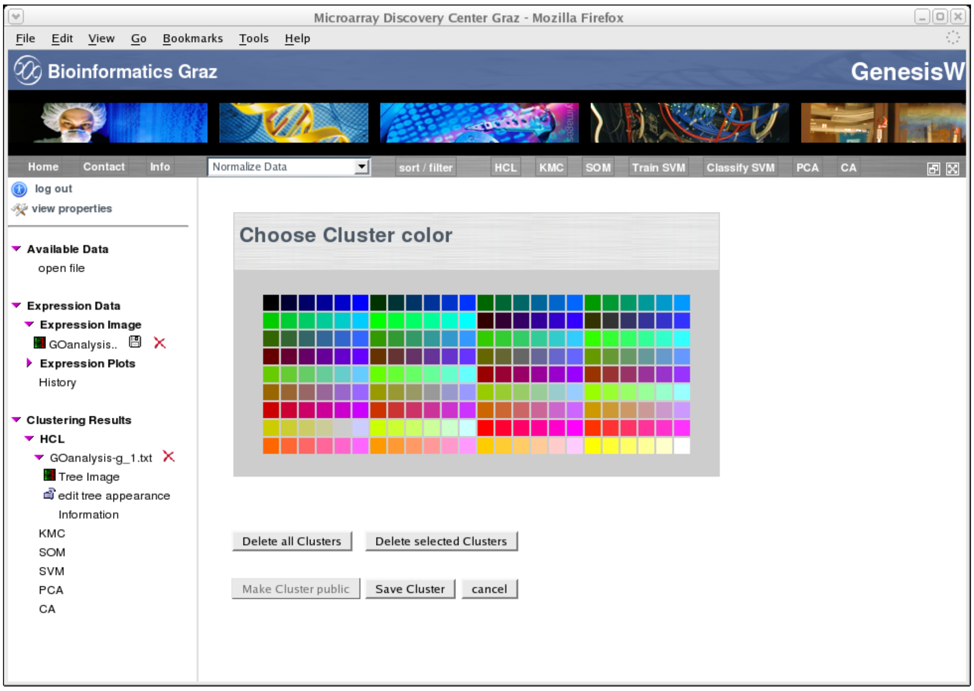


Figure 6.4: Once a cluster has been selected the color can be chosen that should be used to mark this cluster.

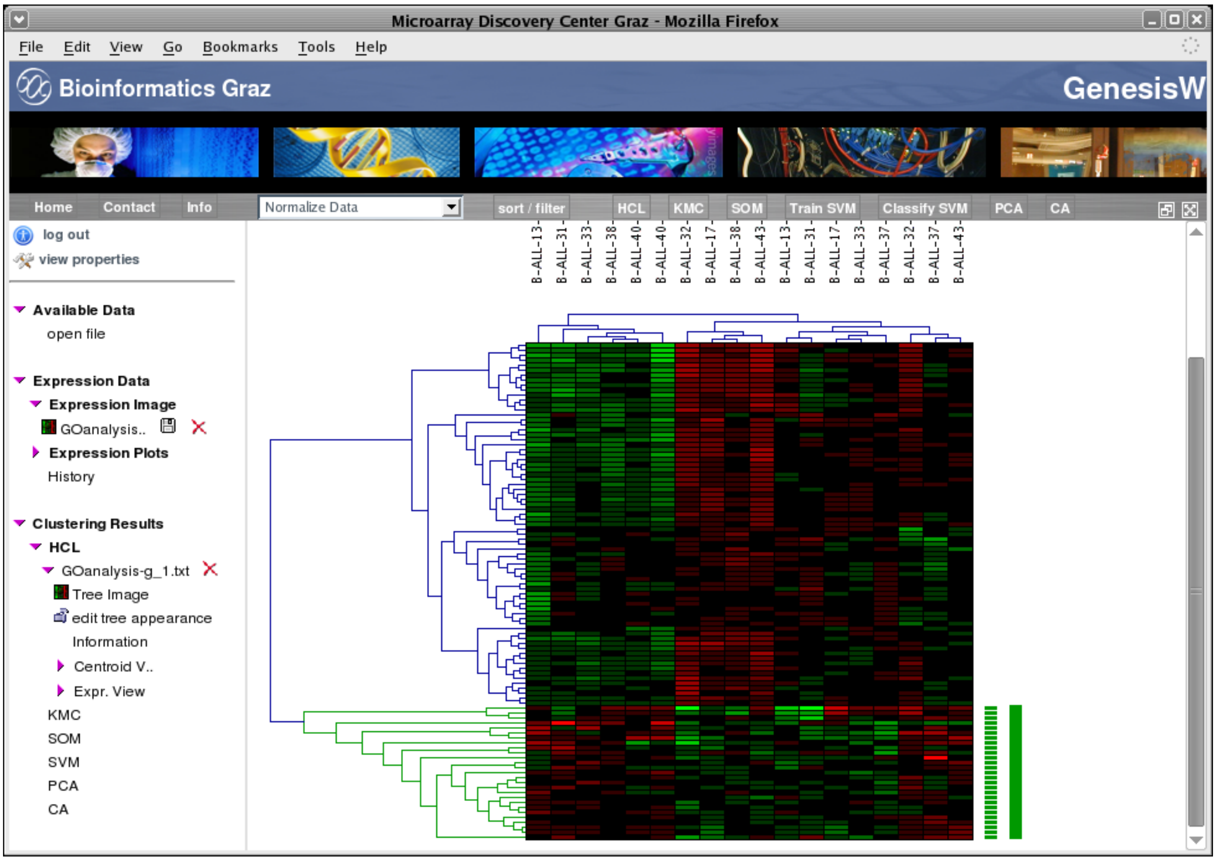
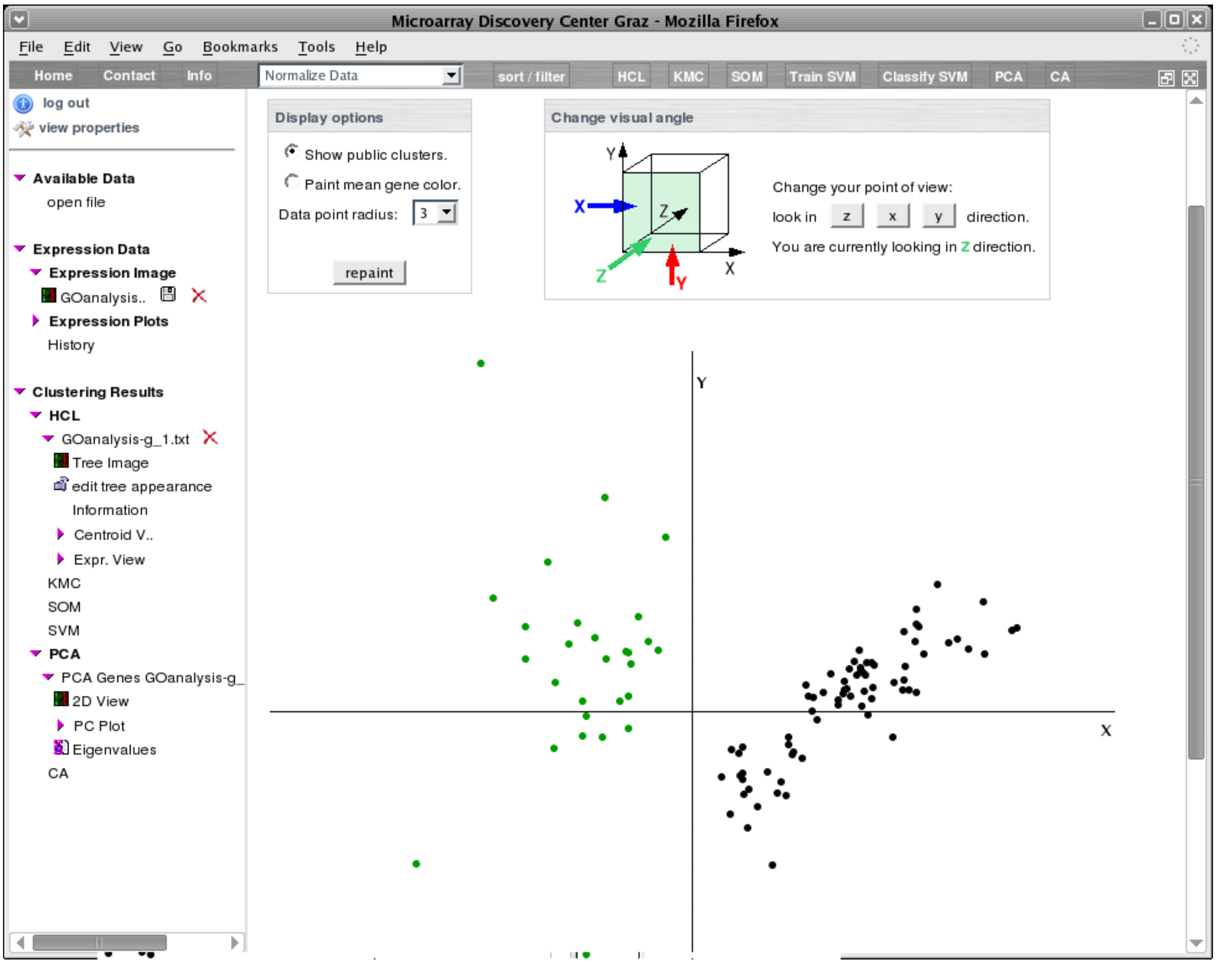


Figure 6.5: HCL cluster result of the test dataset. One cluster has been labeled with a green color.



● Show public clusters.
○ Paint mean gene color.
Data point radius: 3

Change your point of view:

You are currently looking in Z direction.

Figure 6.6: PCA cluster analysis of the test data set. The first two principle components are shown. Using the menu in top of the page it is possible to rotate the image. The genes labeled with the green color are those genes that have been marked in the HCL result in figure 6.5

Bibliography

[1] The apache struts web application framework.

[2] Y Benjamini and Y Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser.*, 57:289–300, 1995.

[3] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003. Evaluation Studies.

[4] S Dudoit, J P Shaffer, and Boldrick J C. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[5] David Edwards. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, 19(7):825–833, May 2003. Evaluation Studies.

[6] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[7] Wolfgang Huber, Anja von Heydebreck, Holger Sultmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:96–104, 2002. Evaluation Studies.

[8] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 31(4):e15, Feb 2003.

[9] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.

[10] C Li and W H Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–36, Jan 2001.

[11] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[12] G K Smyth and T P Speed. Normalization of cDNA microarray data. *Methods*, 31:265–273, 2003.

[13] Gordon K Smyth. Limma: linear models for microarray data. In R C Gentleman, V J Carey, S Dudoit, R A Irizarry, and W Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, page Chapter 23. Springer, New York, 2005.

[14] Alexander Sturn, Bernhard Mlecnik, Roland Pieler, Johannes Rainer, Thomas Truskaller, and Zlatko Trajanoski. Client-server environment for high-performance gene expression data analysis. *Bioinformatics*, 19(6):772–773, Apr 2003.

[15] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, Apr 2001.

[16] P H Westfall and SS Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York, 1993.