Manual for arrayQCplot 2.0

1 Introduction

arrayQCplot is a software for the exploratory analysis of microarray data focusing on quality control. It helps biologists check the quality of their data visually and easily. This software is built on R and provides a user-friendly graphical interface for the graphics and statistical analysis. Therefore, naïve users can use arrayQCplot to check their data just by clicking mouse button.

arrayQCplot is a tool for exploring microarray data, especially two channel cDNA microarray data. It is developed to help biologists investigate microarray data in the first step of analysis. For two channel microarray data, arrayQCplot provides the normalization using lowess method, examine the reliability of experiments and check the reproducibility of replicates if available. For the other types of microarray data, for example, one channel cDNA microarray data, Affy data, etc., the limited service is provided with the normalized data.

2 How to install

arrayQCplot is built on the open-source statistical software, R, and provides a user-friendly interface. To use arrayQCplot package, you need to install

- R
  - a free software environment for statistical computing and graphics
  - available at http://www.r-project.org/
- GTK+
  - The premier open-source graphical user interface toolkit.
  - available at http://www.gtk.org
- RGtk2 package
  - allows one to create GTK+ and Gnome widgets from within R, using R functions and expressions
  - available at http://www.ggobi.org/rgtk2/
- cairoDevice R package
  - help create embedded graphics devices within GUIs
  - available at http://www.ggobi.org/rgtk2/
- LPE package library
  - originally developed to do significance analysis of microarray data with small number of replicates

  - we use this package to estimate the lpe variance within treatments
  - available at http://www.bioconductor.org/packages/bioc/stable/src/contrib/html/LPE.html

- arrayQCplot version 2.0
  - available at http://biostats.snu.ac.kr/ graceslee/arrayQCplot/arrayQCplot.html

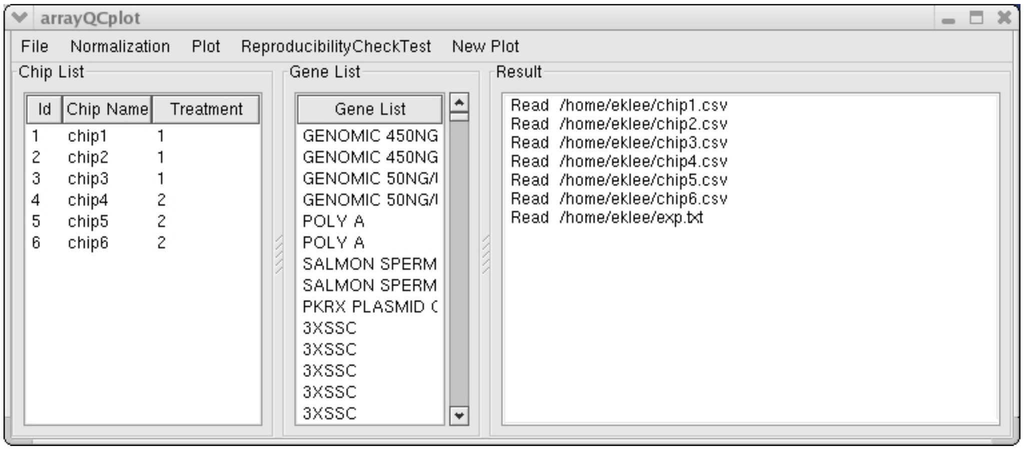
3 How to use arrayQCplot

- Start R
- Load the arrayQCplot library
  `>library(arrayQCplot)`
- Open arrayQCplot
  `>arrayQC.plot()`

3.1 Main GUI

The main GUI of arrayQCplot has the following parts :

- `Chip List` table shows the information about loaded experiments including chip name, and treatments. From this table, the user can select chips that he/she wants to analyze.
- `Gene List` shows the names of loaded genes.
- `Result` table is used to write the result of statistical analysis. In this table, arrayQCplot points out what the user did. All the result from this table can be saved as a file.



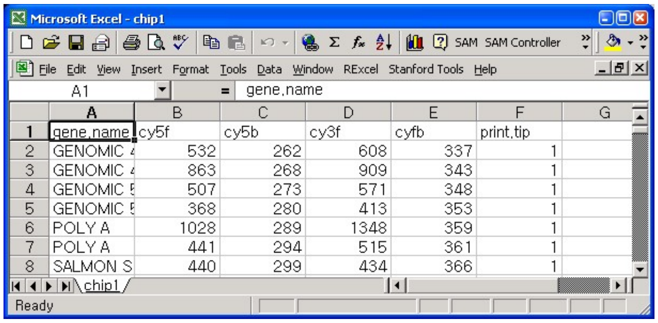
Id	Chip Name	Treatment
1	chip1	1
2	chip2	1
3	chip3	1
4	chip4	2
5	chip5	2
6	chip6	2

Gene List
GENOMIC 450NG
GENOMIC 450NG
GENOMIC 50NG/I
GENOMIC 50NG/I
POLY A
POLY A
SALMON SPERM
SALMON SPERM
PKRX PLASMID (
3XSSC
3XSSC
3XSSC
3XSSC
3XSSC

```
Read /home/eklee/chip1.csv
Read /home/eklee/chip2.csv
Read /home/eklee/chip3.csv
Read /home/eklee/chip4.csv
Read /home/eklee/chip5.csv
Read /home/eklee/chip6.csv
Read /home/eklee/exp.txt
```

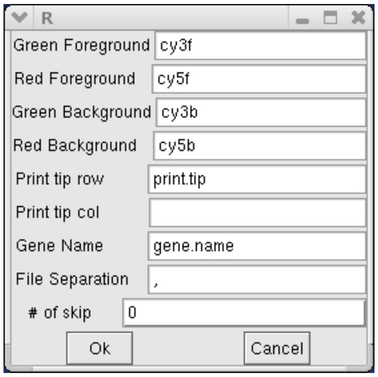
3.2 File

- **Read cDNA Raw data**
  To read two channel cDNA raw data, you need to save your data as following.
  - one chip should be in one file
  - In one file, arrayQCplot needs foreground intensities of cy3 and cy5, background intensities of cy3 and cy5, print tip information, and gene names.
  - recommendation : use Excel spread sheet and save your data as CSV(Comma delimited)

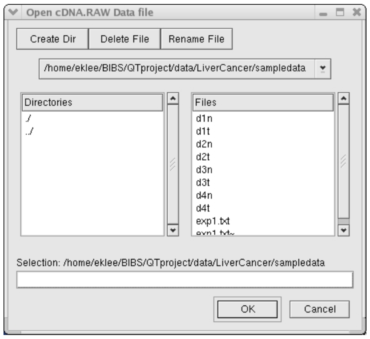


A1 = gene.name

	A	B	C	D	E	F	G
1	gene.name	cy5f	cy5b	cy3f	cyfb	print.tip	
2	GENOMIC 4	532	262	608	337	1	
3	GENOMIC 4	863	268	909	343	1	
4	GENOMIC 5	507	273	571	348	1	
5	GENOMIC 5	368	280	413	353	1	
6	POLY A	1028	289	1348	359	1	
7	POLY A	441	294	515	361	1	
8	SALMON S	440	299	434	366	1	



Green Foreground: cy3f
Red Foreground: cy5f
Green Background: cy3b
Red Background: cy5b
Print tip row: print.tip
Print tip col:
Gene Name: gene.name
File Separation: ,
# of skip: 0



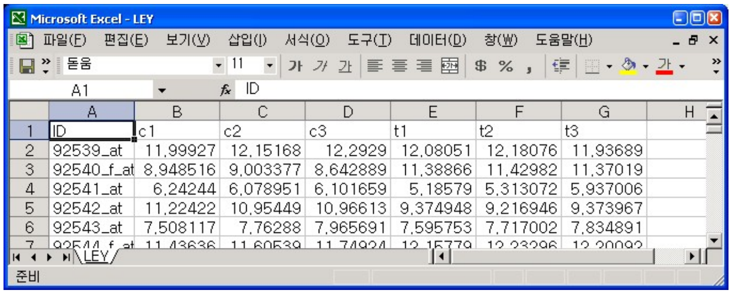
/home/eklee/BIBS/QTproject/data/LiverCancer/sampledata

Directories: ./ ../

Files: d1n d1t d2n d2t d3n d3t d4n d4t exp1.txt exp1.txt~

Selection: /home/eklee/BIBS/QTproject/data/LiverCancer/sampledata

- **Read cDNA Normalized data**
  To read the normalized data, use this menu. In one file, the user needs to input gene names, normalized intensities for each chip.



Microsoft Excel - LEY
파일(F) 편집(E) 보기(V) 삽입(I) 서식(O) 도구(T) 데이터(D) 창(W) 도움말(H)
돋움
11
A1
ID
LEY
준비

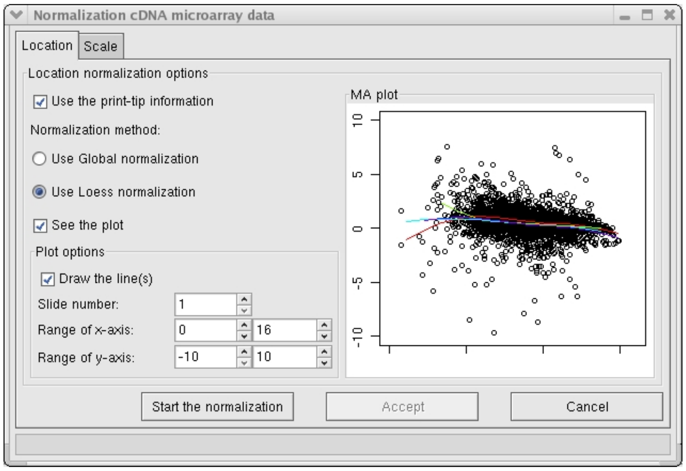
	A	B	C	D	E	F	G	H
1	ID	c1	c2	c3	t1	t2	t3	
2	92539_at	11,99927	12,15168	12,2929	12,08051	12,18076	11,93689	
3	92540_f_at	8,948516	9,003377	8,642889	11,38866	11,42982	11,37019	
4	92541_at	6,24244	6,078951	6,101659	5,18579	5,313072	5,937006	
5	92542_at	11,22422	10,95449	10,96613	9,374948	9,216946	9,373967	
6	92543_at	7,508117	7,76288	7,965691	7,595753	7,717002	7,834891	
7	92544_f_at	11,43636	11,60539	11,74924	12,15779	12,23296	12,20092	

- **Read Exp.Info**
  Some of arrayQCplot functions need experimental information, especially treatment information.

- **Save Result**
  arrayQCplot can save all the results on the Result table.

3.3 Normalization

If the cDNA raw data is loaded, the user needs to normalize data before further analysis. arrayQCplot provides location and scale normalization separately. After Location normalization, the user can select Scale normalization. After all normalization procedure, the user needs to select Accept button. arrayQCplot provides different GUI for normalization. MA plots and box plots are also provided in the Normalization GUI.



Location normalization options

☑ Use the print-tip information

Normalization method:

○ Use Global normalization

● Use Loess normalization

☑ See the plot

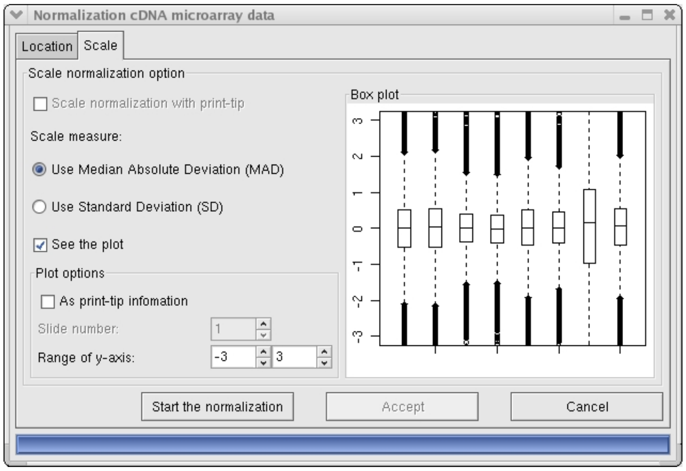
Plot options

☑ Draw the line(s)

Slide number: 1

Range of x-axis: 0 16

Range of y-axis: -10 10



Normalization cDNA microarray data

Location | Scale

Scale normalization option

□ Scale normalization with print-tip

Scale measure:

● Use Median Absolute Deviation (MAD)

○ Use Standard Deviation (SD)

☑ See the plot

Plot options

□ As print-tip infomation

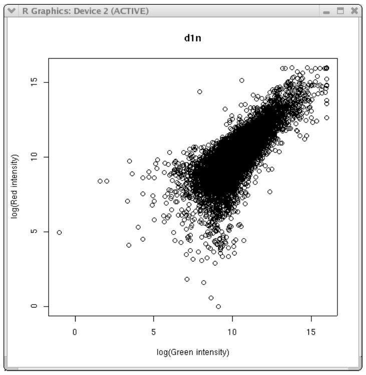
Slide number: 1

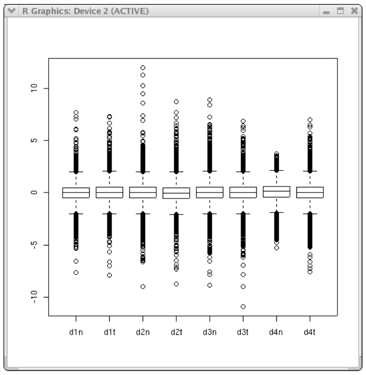
Range of y-axis: -3 3

3.4 Plot

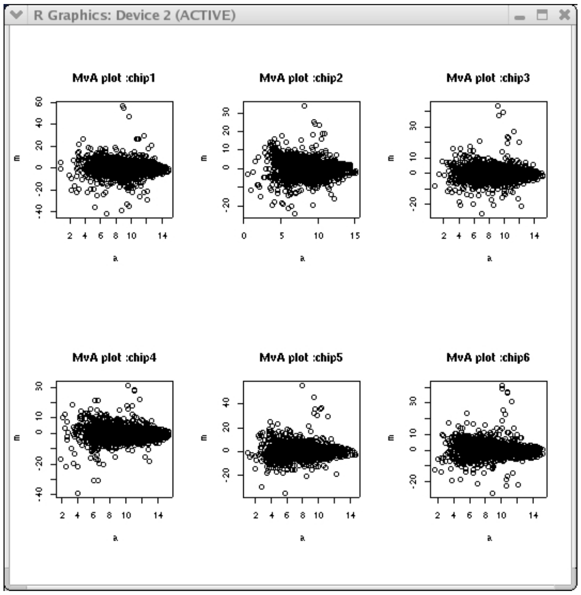
For the normalized data, arrayQCplot provides a variety of basic plots to explore microarray data.

- **Red vs. Green plot**
- **Box plot**

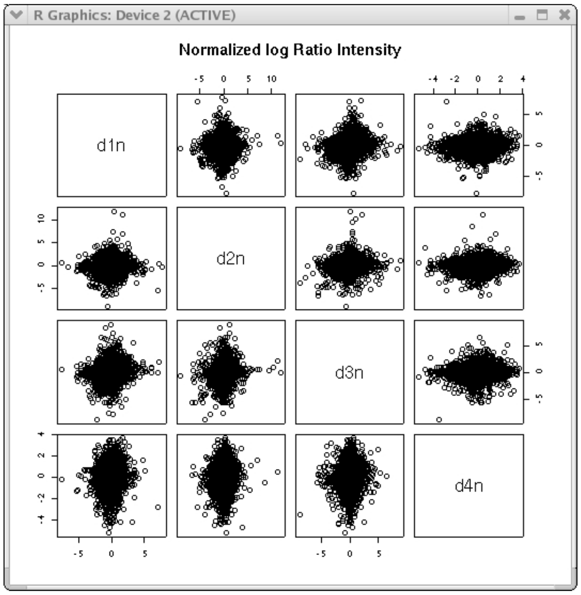




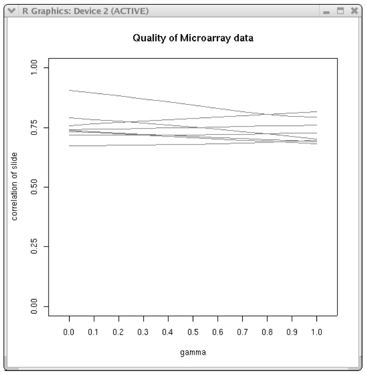
- **MvA plot**

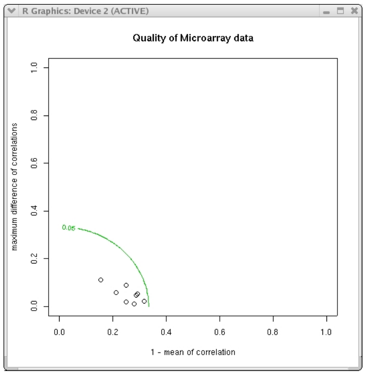


- **Scatter plot matrix : Red/Green/Ratio**

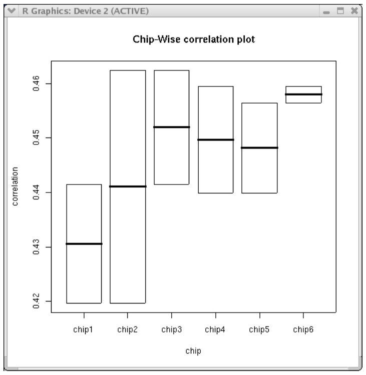


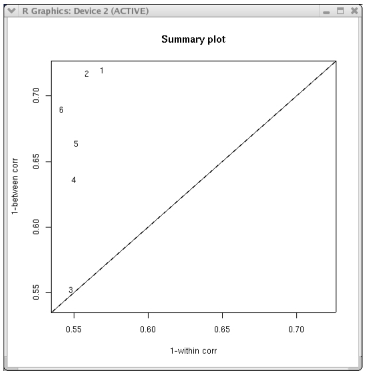
- **Quality Control plot : correlation plot/diagnostic plot**
  To detect outlying chips, arrayQCplot provides two plots, correlation plot and diagnostic plot. For more detailed explanation, see Park et al.(2005).





- **Chip-wise correlation plot : Summary correlation plot**
  To check the reproducibility and quality of experiments, arrayQCplot provides two plots, chip-wise correlation plot and summary correlation plot. For more detailed explanation, see Lee et al.(2006).





3.5 Reproducibility Check Test

Reproducibility Check Test menu in arrayQCplot focuses on the reproducibility, sensitivity and specificity. Therefore, to use this menu, the experiment should have at least two replicates for each treatment level. Let

Y_{ijg} : the normalized log intensity of the jth replicates of the ith treatment for gene g

$i = 1, 2, \cdots, I, \quad j = 1, 2, \cdots, n_i, \quad \text{and} \quad g = 1, 2, \cdots, G$

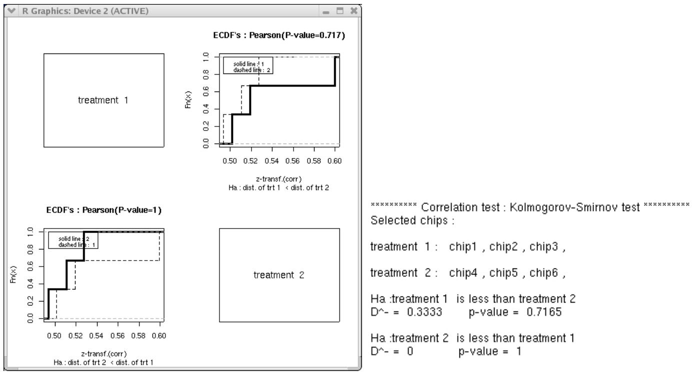
$Y_{ij} = [Y_{ij1}, Y_{ij2}, \cdots, Y_{ijG}]^T$

$r_{ij,kl} = corr(Y_{ij}, Y_{kl})$

$\mathbf{R}_i^w = \{r_{ij,il}, \forall l \neq j\}$

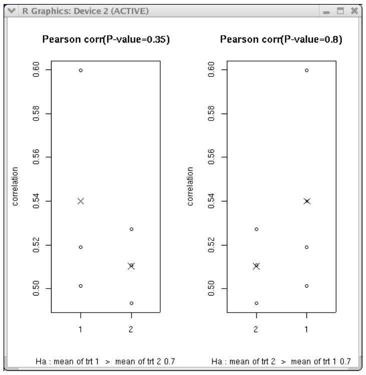
- **Test based on correlation**

1. **K-S test with z-transformation**
   To compare the distributions of correlation sets of two treatments, $\mathbf{R}_i^w$ and $\mathbf{R}_{i'}^w$, we use Kolmogorov-Smirnov test after z-transformation. If $p - value$ is small, $\mathbf{R}_i^w$ and $\mathbf{R}_{i'}^w$ have different distributions. If the empirical CDF's of $\mathbf{R}_i^w$ is lower than the empirical CDF's of $\mathbf{R}_{i'}^w$, the reproducibility and sensitivity of treatment i is better than treatment i'.



```
********** Correlation test : Kolmogorov-Smirnov test **********
Selected chips :

treatment  1 :   chip1 , chip2 , chip3 ,

treatment  2 :   chip4 , chip5 , chip6 ,

Ha :treatment 1  is less than treatment 2
D^- =  0.3333       p-value =  0.7165

Ha :treatment 2  is less than treatment 1
D^- =  0            p-value =  1
```

2. **mean difference test**
   To test the differences of correlation sets of two treatments, $\mathbf{R}_i^w$ and $\mathbf{R}_{i'}^w$, we use Wilcox rank sum test. If $p - value$ is small, $\mathbf{R}_i^w$ and $\mathbf{R}_{i'}^w$ are significantly different. If the mean of $\mathbf{R}_i^w$ is greater than the mean of $\mathbf{R}_{i'}^w$, the reproducibility and sensitivity of treatment i is better than treatment i'.

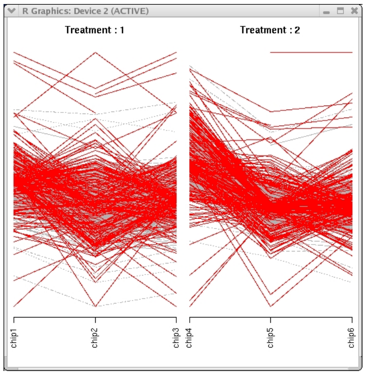

R Graphics: Device 2 (ACTIVE)
Pearson corr(P-value=0.35)
Pearson corr(P-value=0.8)
correlation
0.50
0.52
0.54
0.56
0.58
0.60
1
2
Ha : mean of trt 1 > mean of trt 2 0.7
Ha : mean of trt 2 > mean of trt 1 0.7

```
********** Correlation test : Wilcoxon Rank Sum test **********
Selected chips :

treatment  1 :   chip1 , chip2 , chip3 ,

treatment  2 :   chip4 , chip5 , chip6 ,

Ha : mean of treatment 1  is greater than mean of treatment 2
W =  6          p-value =  0.35

Ha : mean of treatment 2  is greater than mean of treatment 1
W =  3          p-value =  0.8
```

- **Test based on intensities**

  Within treatment
  For gene j, we test whether the mean intensities within treatment are same or not. We assume that for gene g, $Y_{ijg} \sim N(\mu_{ijg}, \sigma_{ig}^2)$.

  Test for each treatment : To check the reproducibility of the treatment i, we test $H_0 : \mu_{i1g} = \mu_{i2g} = \cdots = \mu_{in_ig}$. After applying step 1 in the overall test, $\quad \sum_j \left(\frac{Y_{ijg}-\bar{Y}_{i.g}}{\hat{\sigma}_i}\right)^2 \sim \chi^2(n_i - 1)$.



```
***** Differently expressed Gene : within treatment *****
*****                  use Intensity                   *****

Selected chips :   chip1  chip2  chip3  chip4  chip5  chip6
   (delete 679 genes due to missing)

Test for each treatment
 Treatment : 1
    Concordant gene : 6803
    Discordant gene : 198
    gamma           : 0.9434
 Treatment : 2
    Concordant gene : 6791
    Discordant gene : 210
    gamma           : 0.94
```

4 References

Park, T., Yi, S-G., Lee, S. Y., and Lee, J. K. (2005) Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. *Biotechniques 38:463-471*
Lee, E. K., Yi, S-G, and Park, T. (2006) Exploratory Methods for Checking Quality of Microarray data. *Technical report, BIBS, Seoul National University*