2012
2013

FACULTY OF SCIENCES

Master of Statistics: Bioinformatics

Masterproef

Diagnostic tools for biclustering output

Promotor :
Prof. dr. Ziv SHKEDY

Mengsteab Fantahu Aregay

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt

KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2012
2013

FACULTY OF SCIENCES

Master of Statistics: Bioinformatics

Masterproef

Diagnostic tools for biclustering output

Promotor :
Prof. dr. Ziv SHKEDY

Mengsteab Fantahu Aregay

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

universiteit hasselt | Maastricht University

Diagnostic tools for biclustering output

The R package 'BcDiag'

Aregay Mengsteab F.

Supervisors:

Prof. Dr. Ziv SHKEDY

Mrs. Tatsiana KHAMIAKOVA

Mr. Martin OTAVA

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics.

2010-2013

Acknowledgements

First of all I would like to thank you God for always being with me and for an endless love and wisdom you put on my heart.

My sincere gratitude goes to my Supervisor at Univesiteit Hasselt, Belgium, Prof. Dr. Ziv SHKEDY for his guidance and support for the development of this software package. I also extended my gratitude to Mrs. Tatsiana KHAMIAKOVA and Mr. Martin OTAVA, for giving me supportive materials and for your supportive comments.

To the entire staff of Centre for Statistics, Universiteit Hasselt, I thank you for all the support you gave me in the course of my study in Belgium.

Finally, I deeply thank you for my best friends and relatives for funding my studies which I would have not been able to pursue a master degree.

Lists of abbreviations and definitions

Functions:

exploreBic: a function used to plot both (biclustered and non biclust) summary statistics

exploreCalc: a supportive function to exploreBic

explorePlot: a supportive function for exploreBic

exploreOnlybic: a function used to plot summary statistics only for biclustered data.

indexBic: a function used to extract indices of the biclustered matrix from the dataset.

profileBic: a function used to plot profile biclusters

profilePlot: a supportive function to profileBic function.

profileAll: a supportive function for profileBic function.

plotOnlybic: a supportive function to exploreOnlybic.

writeBic: a function used to write the biclust results to a text file.

Frequently used Parameters:

***dset*:** dataset

***bres*:** biclustered result

***mname*:** method name (FABIA, ISA or biclust)

***bnum*:** bicluster number

Table of Contents

Abstract

A BcDiag is a biclustered R package which provides diagnostics and visualization plots for biclustered and clustered data matrix. Five main functions and five supportive functions are included in the package. Three different biclustered packages have made a link to create the bicluster output. More than 8 bicluster methods has supported by the package to extract the biclustered output. It provides profile plots such as 3D, histogram, box plot and parallel plots to visualize the biclustered and cluster out come. Additional diagnostic plots are included to visualize the mean, median, variance, mad and quantile summary statistics of biclustered and clustered data. Additionally, the package presented residual plots for biclustered data based on the outcome of mean and median polish of two ways ANOVA model. Finally, the package has extended the function writebic from biclust package, to hold up additional algorithms from fabia and isa packages

1 Introduction

Gene expression microarray technology has become a central tool in biological research. Microarrays measure the mRNA levels of thousands of genes, perhaps all genes of an organism, within a number of different samples (conditions), in a single experiment while traditional methods work on a "one gene in one experiment" basis (Ben-Dor et al, 2003) The samples may correspond to different time points or different environmental conditions. In other cases, the samples may have come from different organs, from cancerous or healthy tissues, or even from different individuals. Simply visualizing this kind of data, which is widely called gene expression data or simply expression data, is challenging and extracting biologically relevant knowledge is harder still.

The analysis of microarray data poses a large number of exploratory statistical aspects including clustering and biclustering algorithms, which help to identify similar patterns in gene expression data and group genes and conditions into subsets that share biologically significance. The goal of clustering algorithms is to partition the elements (genes) into sets, or clusters, while attempting to optimize homogeneity (elements inside a cluster are highly similar to each other) and separation (elements from different clusters have low similarity to each other). Clustering methods can be applied to either the rows or the columns of the data matrix, separately (Sharan et al., 2002). Biclustering methods, on the other hand, perform clustering in the two dimensions simultaneously. This means that clustering methods derive a global model while biclustering algorithms produce a local model. When clustering algorithms are used, each gene in a given gene cluster is defined using all the conditions. Similarly, each condition in a condition cluster is characterized by the activity of all the genes. However, each gene in a bicluster is selected using only a subset of the conditions and each condition in a bicluster is selected using only a subset of the genes. The goal of biclustering techniques is thus to identify subgroups of genes and subgroups of conditions, by performing simultaneous clustering of both rows and columns of the gene expression matrix, instead of clustering these two dimensions separately (Tanay et al., 2002).

Several authors have proposed different biclustering algorithms, each of which has strengths and weaknesses for the application in different biological scenarios (Madeira and Oliveira, 2004). A comparative study has shown that there are significant differences in performance among biclustering approaches, depending on the biological problem that is examined (Prelic

et al., 2006). Since every algorithm is subject to a specific mathematical problem formulation, it cannot be expected that a single approach is well suited for all scenarios. Accordingly, it can be useful in practice to try different approaches and to choose that algorithm that delivers the best results. Although implementations are available for some of the proposed biclustering algorithms, each program may be accompanied by a different user interface and use different input and output formats, which in turn makes the application of several methods a time-consuming task.

It is important to identify cluster and bicluster data. For this reason, Figure 2 explained the difference between cluster, bicluster and non cluster data. A cluster is when a data matrix has similar characteristics shared in common either by rows or columns (Figure 2-a). But if a data has some similarity measured in both rows and columns simultaneously, we call it bicluster (Figure 2-b). When two biclusters have something in common we call them overlapped biclusters. Hence, Overlapped biclusters are a special type of biclusters.

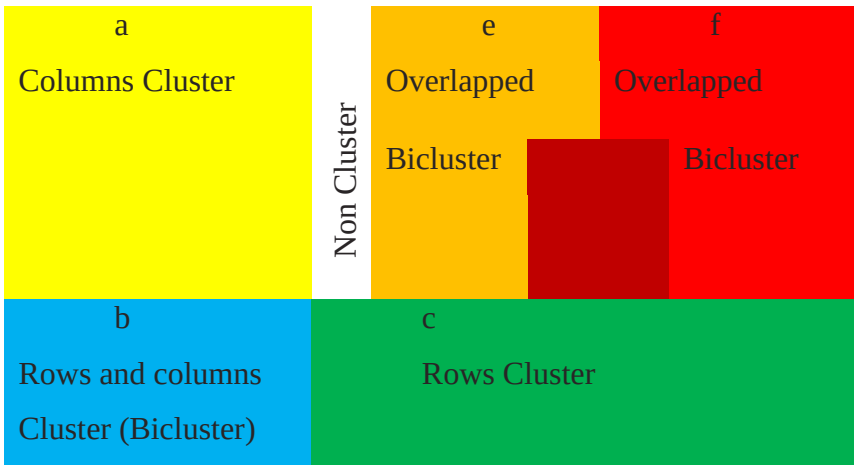


Figure2: *The Cluster, Bicluster Overlapped Clusters and Non clusters data matrix*

There are many Bicluster algorithms existed for different types of bicluster methods. Some of them are discussed in this report under section 3. Even though developing bicluster algorithms are increasing, visualization of the computed biclusters still remains an open issue. In this paper, we present a bicluster visualization R package for profiling and summarizing the gene expression levels. As it has shown in Figure 1, we have two target data to visualize. 1) Clustered columns vs. Bicluster (clustered rows and columns). 2) Clustered rows vs. Bicluster (clustered rows and columns). We used different plots such as box plot,

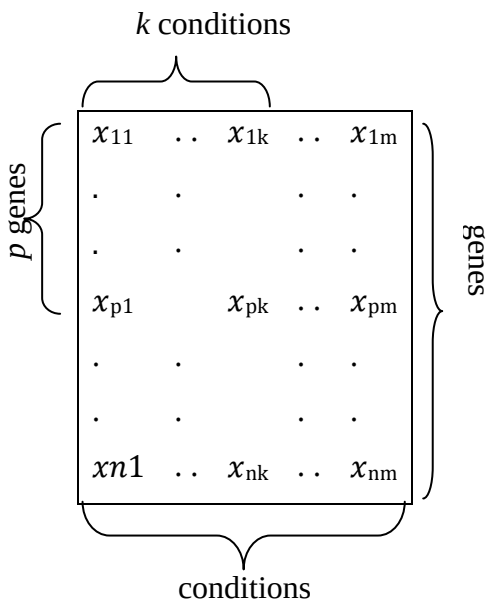
histogram, profile plot and 3D plot to visualize the profile, to diagnostic the target data based on their mean, variance, median, mad and quantile. Moreover, we develop a function to write the bicluster results from different algorithms in a text format. We used bicluster results obtained from Factor Analysis for Bicluster Acquisition (FABIA), Iterative Search Algorithm (ISA), and from all algorithms of biclust package as an input data to our application. Additionally, our application has options to the user to choose different plots to visualize the data. The residual plots will help us to diagnostic possible outliers, normality and constant variance of biclustered data. We used R software to develop the Package.

This report is structured as follows. Section 2 devoted to the description of the data. Section 3 discussed the different biclustering algorithm. Section 4 will be explained about a software development. In Section 5, the results will be presented. Finally, a discussion and conclusion will be drawn in Section 6.

2 Data

In this report, we analyzed two data sets; Breast cancer and diffuse large B-cell lymphoma. Breast Cancer data were discussed in detail in Van't Veer *et al*. 2002, where the goal was to find a gene signature to predict the outcome of a breast cancer therapy. There are 97 samples and 1213 genes. The second data set which is large B-cell lymphoma has 180 samples and 661 genes. The objective was to predict the survival after chemotherapy and to identify these subclasses directly by biclustering.

Table 1: *general structure of gene expression data matrix*



The aim of a biclustering analysis is to find a sub matrix for which a group of p genes expressed on a group of k conditions.

3 Bicluster algorithms

Different applications have their own of identifying biclusters. The most common types of bicluster classes are; 1) Biclusters with constant values (Figure 1-A), 2) Bicluster on constant values with rows or columns (Figure 1-B), 3) Bicluster coherent values (Figure 1-C) and 4) Bicluster coherent evolutions (Figure 1 D) (Madeira and Oliveira, 2004).

A		
1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	1.0

B		
1.0	2.0	3.0
1.0	2.0	3.0
1.0	2.0	3.0

C		
1.0	2.0	5.0
2.0	3.0	6.0
4.0	5.0	8.0

D		
70	13	19
29	40	49
40	20	27

Figure 1: *Different types of Bicluster classes*

Bicluster methods have become more popular in different fields of two way data analysis and a wide variety of algorithms and analysis methods have been proposed. In this section, we will describe some of the methods for the analysis of biclustering.

3.1 Plaid Model

These are a form of two-sided cluster analysis that allows clusters to overlap Plaid models. It incorporates additive two way ANOVA models within the two-sided clusters. The motivating application is the search for interpretable biological structure in gene expression microarray

data. The plaid model allows a gene to be in more than one cluster or in none at all. It also allows a cluster of genes to be defined with respect to only a subset of samples, not necessarily with respect to all of them (Lazzeroni *et al.*, 2002).

3.2 FABIA

FABIA: (Factor Analysis for Bicluster Acquisition) is a multiplicative model, which accounts for linear dependencies between gene expression and conditions. It also captures realistic non-Gaussian data distributions with heavy tails as observed in gene expression measurements. Moreover, it is a model-based technique for biclustering that is clustering rows and columns at the same time. In FABIA, biclusters are found by factor analysis where both the factors and the loading matrix are sparse. FABIA utilizes well understood model selection techniques like the EM algorithm and variation approaches and is embedded into a Bayesian framework. FABIA ranks biclusters according to their information content and separates spurious biclusters from true biclusters (Hochreiter *et al*., 2010).

3.3 Motif

A conserved gene expression motif is a subset of genes that is simultaneously conserved across a subset of samples. In this method, a gene's expression level is said to be conserved across a subset of samples if the gene is in the same state in each of the samples in this subset. A conserved gene expression motif or xmotif is a subset of genes whose expression is simultaneously conserved for a subset of samples; we say that each of these samples matches the motif. It is a useful and concise representation of gene expression data in the form of conserved gene expression motifs or xmotifs. These motifs capture the degree of conservation in the gene expression profiles of the samples belonging to a class at two levels: (i) each gene in the motif is similarly-expressed in each of the samples and (ii) all the genes in the motif are simultaneously conserved in all these samples. It is believed that this representation has the potential to capture many key biological properties implicitly present in gene expression data (Murali *et al*., 2003).

3.4 Iterative Signature Algorithm (ISA)

The Iterative Signature Algorithm (ISA; Ihmels et al., 2002, Ihmels et al., 2004) is a biclustering method in which the input of a biclustering method is a matrix and the output is a

set of biclusters that fulfill some criteria. A bicluster is a block of the potentially reordered input matrix. In this report, the Iterative Signature Algorithm 2 (isa2) was used.

In this report, we used both median polish and mean for exploratory analysis using residual plots. The primarily advantage of median polish over analysis by means is its resistance to outliers. On the other hand the classical analysis by means have the following advantages; no need of iteration, always produce a unique fit, minimizes the sum of square residuals, easy to understand and explain and easy to program on computer or calculator (Houglin et al., 1983).

4 Structure and Capacity of BcDiag Package

4.1. Over all structure of the package

There have been developed different visualization tools for biclustered algorithms. Some of them are used to visualize single bicluster using parallel plots (Barkow *et al*., 2006, Cheng *et al*., 2007) and others used to visualize multiple biclusters (Santamaria *et al.*,2008). Our application is an alternative approach for single bicluster visualization. It is an R package application that helps: 1) to visualize specific biclustered data and clustered rows/ columns using profile plot simultaneously. 2) To diagnostic the summary statistics of biclustered and clustered rows/ columns such as: mean median, variance and mad. Moreover, the package has multi optional plots such as: 3D plot, line plots, box plot an histograms to profile the data. Additionally, the package supports different biclustered algorithms from three r packages: biclust, isa2 and fabia. Figure 3 explained the New R package structure and the over flow of data, the link between the functions and other packages.

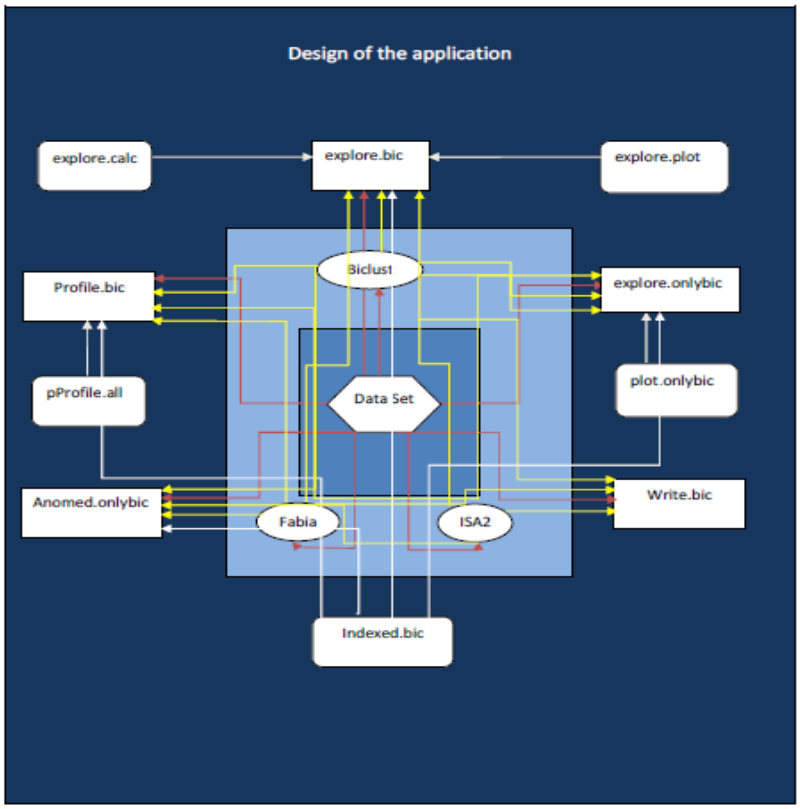


Figure 3*: Design and Structure of BcDiag*

4.2 Capacity of the package

As a major category, we can divide the capacity of the package in two four sub categories. 1) Visualizing profiles plots and diagnostics plots for a specific clustered rows or/ columns and bicluster rows and columns. 2) Diagnostics analysis only for bicluster summaries such as: Mean median, variance and mad. 3) Residual analysis for mean and median polish based on two way ANOVA model. 4) Extending a biclust function called *writeclust* to have more features for other biclust algorithms such as Fabia and isa2. The first three major tasks have explained in Figure 4.

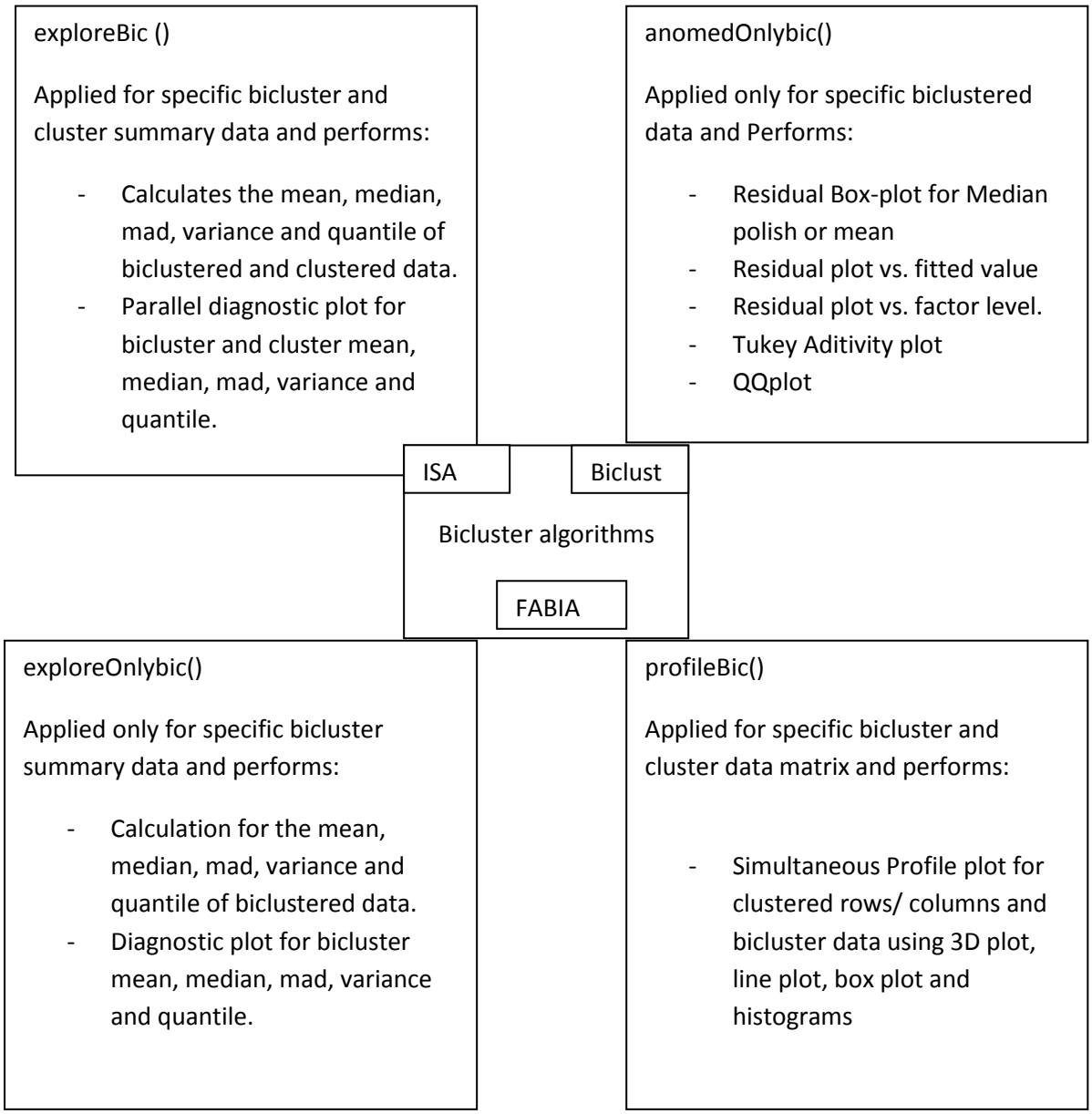


Figure 4: *Summary for the capacity of Major functions of the BcDiag*

4.2.1. Plots for bicluster and cluster data grouped by columns or rows

4.2.1.1. ProfileBic (): The profile plot function

It is a function used to plot profile of individual biclusters. It has mainly four different types of plots: box-plot, line plot, histogram and 3D plot. It has a feature of handling an error and alerting the user about it. Moreover, the function has a choice for the end user to display either a single plot or all plots in a single frame work. It has two supportive functions; ***profileAll*** (used to plot all profile plots based on their genes or conditions) and indexBic

(used to extract the index of bicluster from the data set and returns index values of genes and conditions). Figure 5 was extracted from the following r code.

```
>data(breastc);
>require(fabia); fab<-fabia(breastc); Par(mfrow=c(1,2))
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="lines",gby="genes",t
eta=-30,ph=30)
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="3D",gby="genes",teta
=-30,ph=30)
```

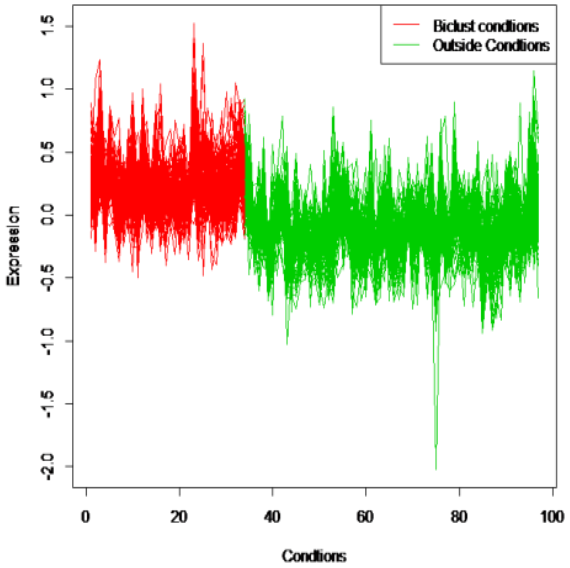


Figure 5.a: *line plot bicluster grouped by conditions*

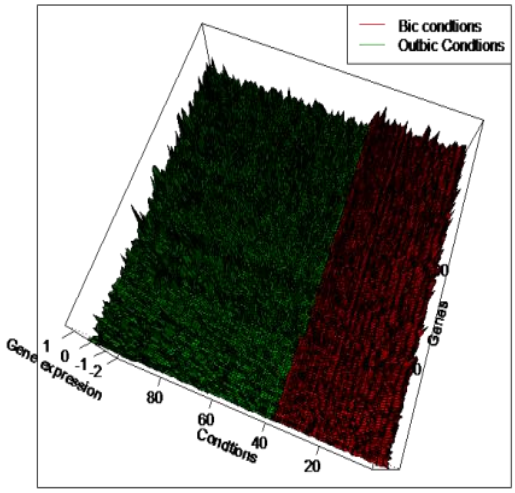


Figure 5.b: *3D bicluster grouped by Conditions*

On the other hand a profile can be plotted using histogram and box plot for the same data. This will increase understanding of the data in different way. Figure 6 was resulted from the following r code from BcDiag package.

```
>data(breastc);
>require(fabia); fab<-fabia(breastc); Par(mfrow=c(1,2))
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="boxplot",gby="genes"
,teta=-30,ph=30)
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="histogram",gby="gene
s",teta=-30,ph=30)
```

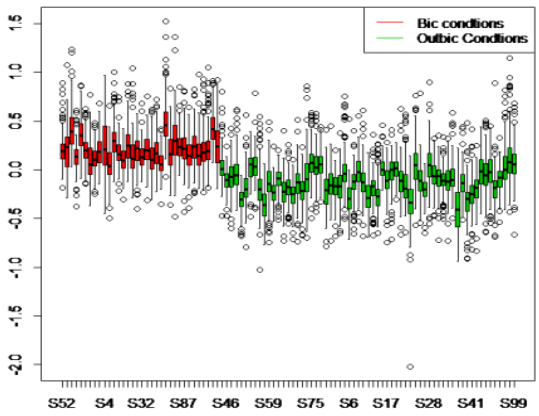


Figure 6.a: *Box plot bicluster grouped by Columns*

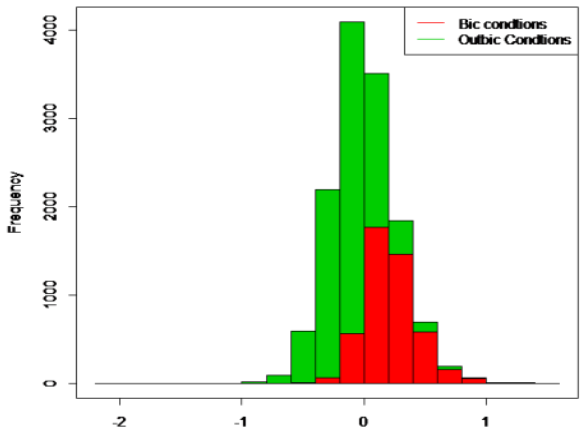


Figure 6.b: *Histogram bicluster grouped by columns*

Profile plots can also be plotted based on their row/genes. Therefore, we need to change the parameter *group by* in to genes/ rows. The rest parameters of *profileBic* function remain unchanged. The following code was taken from the package manual and it provides the 3D and line plot (Figure 7).

```
> data(breast);
>require(fabia); fab<-fabia(breastc)
>Par(mfrow=c(1,2))
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="lines",gby="genes",teta=-30,ph=30)
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="3D",gby="genes",teta=-30,ph=30)
```

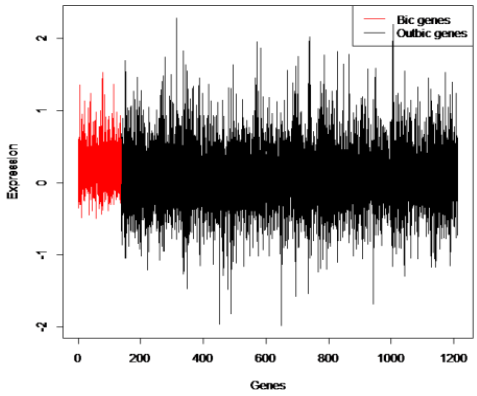


Figure 7.a: *Time plot bicluster vs. cluster data*

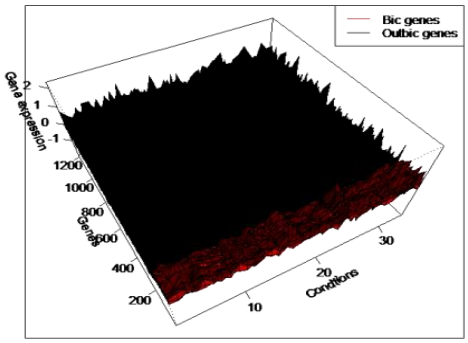


Figure 7.b: *3D bicluster vs. cluster plot*

The same function and parameters can be used to extract the histogram and box plot. but we need to modify again the pfor parameter to be histogram or box-plot. Figure 8 was extracted from the code below.

```
>data(breastc);
>require(fabia); fab<-fabia(breastc)
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="boxplot",gby="genes",teta=-
30,ph=30)
>profileBic(dset=breastc,bres=fab,mname="fabia",bplot="histogram",gby="genes",teta=-
30,ph=30)
```

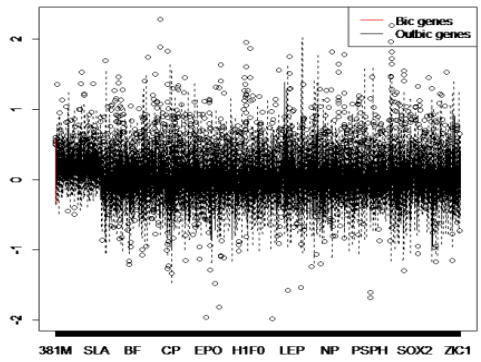


Figure 8.a: *Box-plot bicluster vs. cluster data*

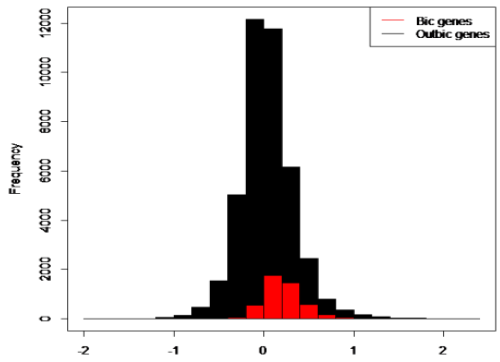


Figure 8.b: *Histogram bicluster vs. cluster data*

4.2.1.2. exploreBic (): The explore clustered rows/columns vs. biclustered data .

These are plots used to visualize the summary of bicluster data such as the mean, the median, the variance, quartile plots and the median absolute deviation of the bicluster (genes and conditions) and the non biclustered genes or conditions. It will compute the summary statistics and plot them in a two groups within a single frame work. It has three supportive functions; ***exploreCalc*** (calculates the summary statistics of each row/column matrix and returns the values) and ***explorePlot*** (plots the profile of all summary statistics based on genes or conditions) and ***indexedBic*** (identifies bicluster from the data set using the indices). Moreover, the function has an error handling procedure. It will alert the user which parameter has to be used for a specific task.

Figure 8 was extracted from the exploreBic function and the biclustered and clustered was grouped based on their columns/conditions. And the output was generated from the R code below.

```
>data(breastc)
>Par(mfrow=c(1,2)
>require(fabia); fab<- fabia(breastc)
>exploreBic(dset=breastc,bres=fab,gby="conditions",pfor="mean",mname="fabia")
>exploreBic(dset=breastc,bres=fab,gby="conditions",pfor="variance",mname="fabia")
```

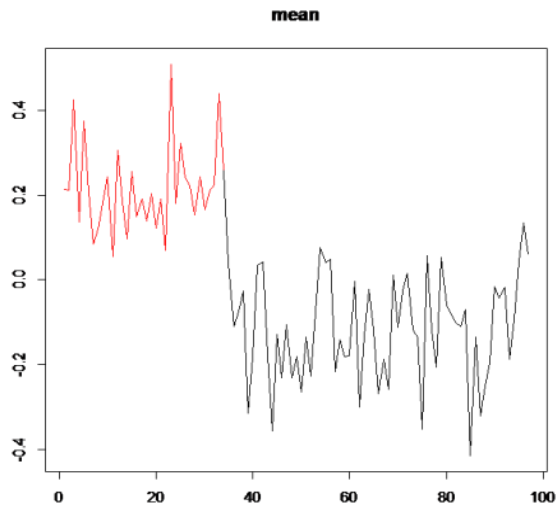


Figure 9: *Mean of biclust vs. cluster by column*

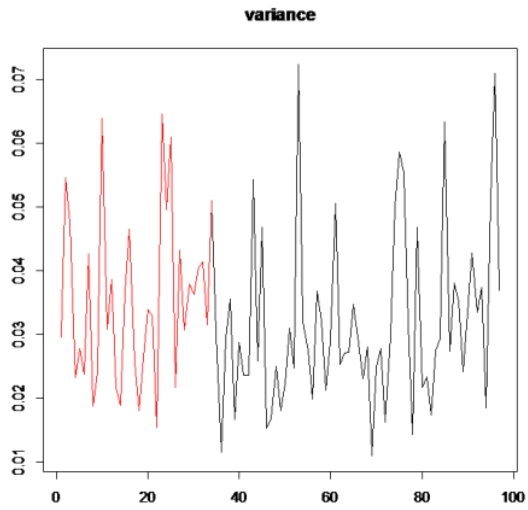


Figure 9.b: *Variance of biclust vs. cluster by column*

Similar plots can be obtained for median and mad (median absolute deviation). Additionally, biclustered data can be also computed based on the rows or genes in our case. Therefore, the exploreBic function can be used to extract the plot using similar parameters except that the group should be changed to genes/rows. The R code below extracted for the mean and variance plot grouped by genes. It has presented on Figure 10.

```
>data(breastc)
>require(fabia); fab<- fabia(breastc)
>exploreBic(dset=breastc,bres=fab,gby="conditions",pfor="all",mname="fabia"
)
```

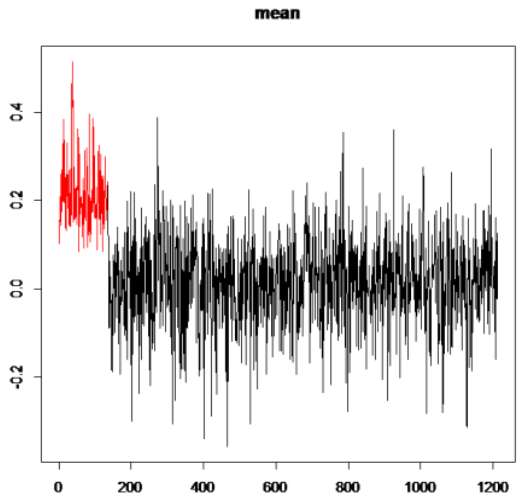


Figure 10.a: *Mean of biclust vs. clusters, grouped by genes.*

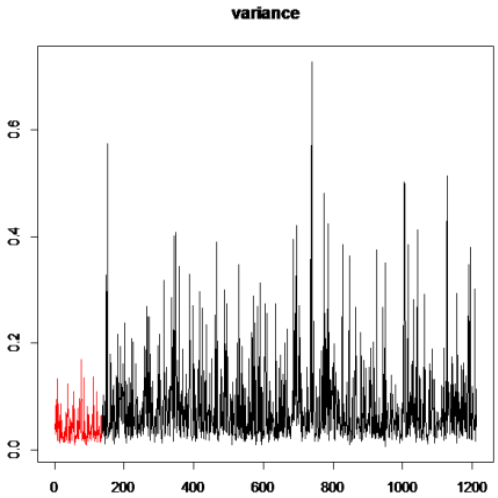


Figure 10.b: *Variance of biclust vs. clusters grouped by genes*

4.2.2. Plots only biclustered data grouped by columns or rows

4.1.2.1. exploreOnlybic (): The explore plot for only bicluster data.

In the previous function (***exploreBic)***, our target data was biclustered (genes and conditions) and non biclustered (genes or conditions). In this case, we are only interested on the biclustered data. Similarly, most of the explanations for ***exploreBic*** are also works for ***explor.onlybic*** function. It has two supportive functions; the ***indexedBic*** and ***plotOnlybic.*** It provides diagnostics analysis plot for mean, variance, median and mad. In this case we will illustrate for mean and variance biclusters grouped by conditions. R codes used to draw the plot in Figure 11 should look like the following.

```
>data(breastc)
>require(fabia; fab<- fabia(breastc); par(mfrow=c(1,2))
>exploreOnlybic(dset=breastc,bres=fab,fit="mean",gby="genes",mname="fabia",bnum=1)
>exploreOnlybic(dset=breastc,bres=fab,fit="variance",gby="genes",mname="fabia",bnum=1)
```

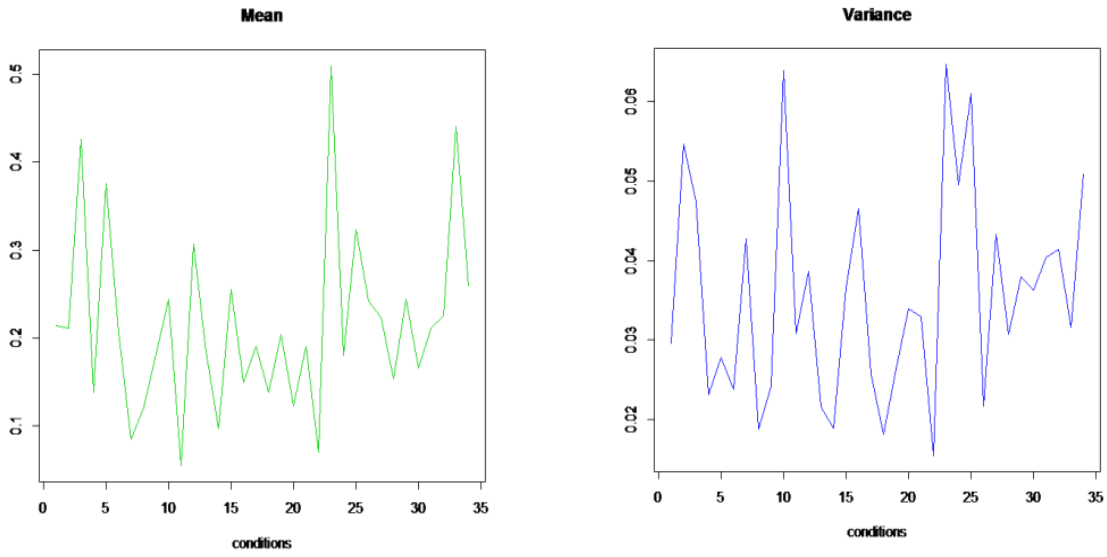


Figure 11: *Biclust mean (**left**) and variance (**right**) by condition*

Another way of diagnostics the summary of bicluster is based on the genes/ rows. The exploreOnlybic function provides the appropriate plot based on the required parameters. From the code below, we can create the mean and variance diagnostics plot grouped by genes/rows for Figure 12. Similar way can also be applied for median and mad diagnostics plot.

```
>data(breastc)
>require(fabia; fab<- fabia(breastc); par(mfrow=c(1,2))
>exploreOnlybic(dset=breastc,bres=fab,fit="mean",gby="genes",mname="fabia",
bnum=1)
>exploreOnlybic(dset=breastc,bres=fab,fit="variance",gby="genes",mname="fab
ia",bnum=1)
```

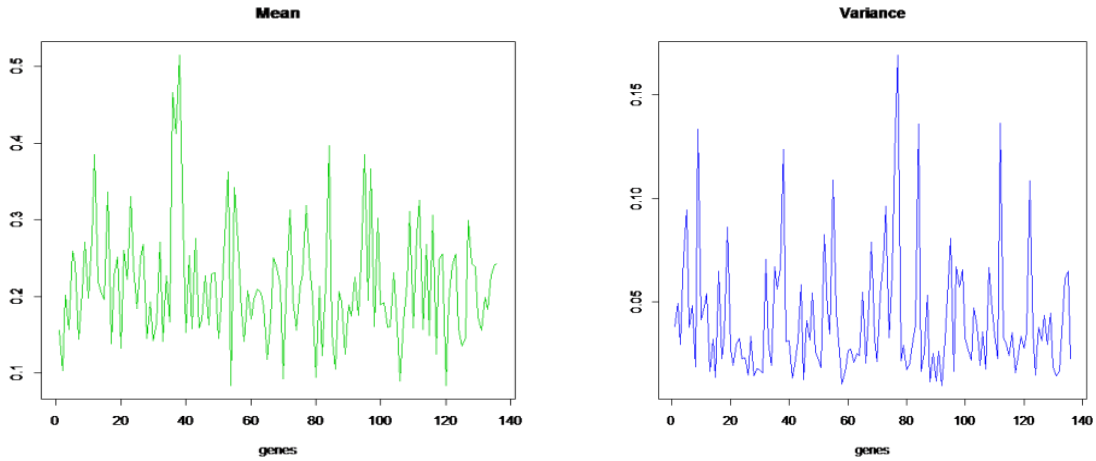


*Figure 12 biclust mean (**left**) and variance (**right**) by condition*

4.1.2.2. AnomedOnlybic(): The residual plot for two way ANOVA model.

It is a residual plot function using mean and median polish of biclustered data. It will obtain the residuals from two way ANOVA model and using Median polish function. It has different residual plots related to the factors (genes and conditions) based on the users choice. It uses box plots and scatter plots to draw the residuals versus the genes and conditions computed using both methods (mean and median polish).

```
>data(breastc)
>require(biclust)
>bic<- biclust(breastc, method=BCPlaid())
>anomedOnlybic(dset=breastc,bres=bic,fit="aplot",mname="biclust")
```

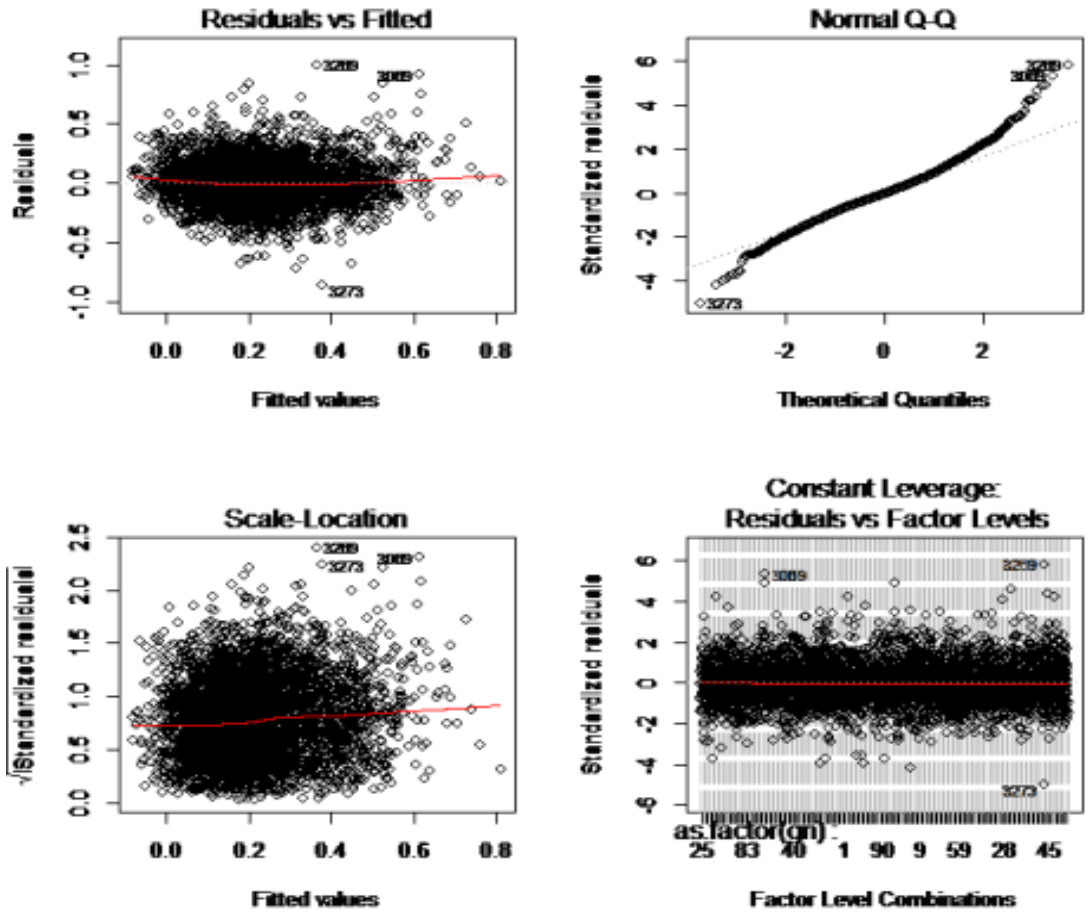


Figure 13: *Diagnostics residual plots for outliers and normality of biclustered data.*

Residuals can be estimated either from the mean or median of blocked effect. The median polish was used to estimate the residuals from the median and to compare with the mean estimation. Figure 16 shows the residual plots obtained from ANOVA. The top box plot in Figure 14 shows the residual plot obtained from ANOVA while the bottom box plot of Figure 14 indicates the residual plot which is obtained from two way table using median polish.

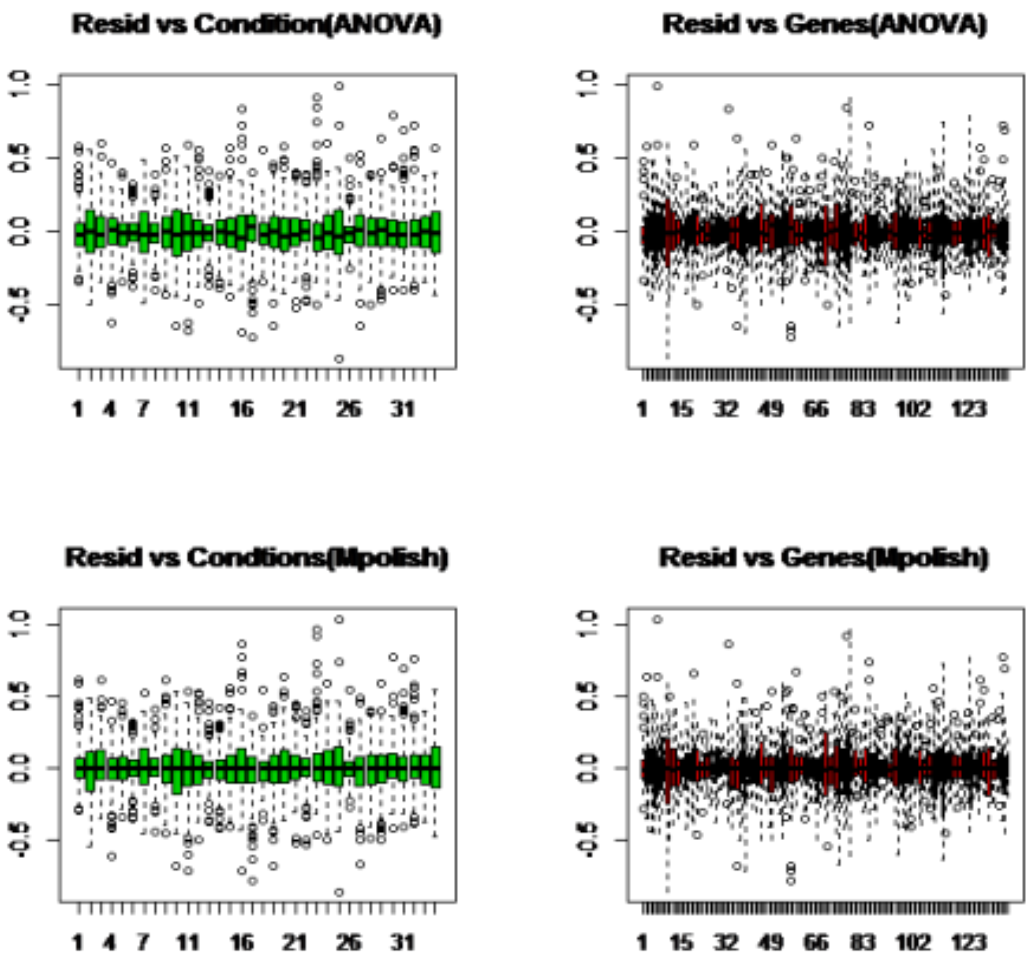


Figure 17: *Diagnostics bicluster residual using box plots based one mean and median polish*

4.1.3. WriteBic (): A function to write bicluster results

Other useful function was developed to write the biclust results that have identified by different algorithms. It will take bicluster results as an input and it will write it to text file. As

a result it will display the total number of biclusters found, the dimension of each biclust and the names of genes and conditions. (Figure 18)

```
>data(breastc)
>require(isa2)
>isa<- isa(breastc)
>writeBic(dset=breastc,fileName="isabreast.txt", bicResult=isa,
bicname="Biclust results for isa", mname="isa2")
```

```
3
Biclust results for plaid model
60 28
ABAT ACADSB ADCY1 AGTR1 ALCAM APBB2 AREG BAG1 BCL2 C10orf116 C16orf45 C4A
CACNA1D CACNG1 CELSR2 CHAD CITED1 COL4A5 COX6C CPB1 CYP2B7P1 CYP4B1
DHRS2 ECM1 EEF1A2 ERBB3 ERBB4 ESR1 FASN FBP1 GATA3 GSTM3 HD HOXB6 HTR7
ITPR1 KIAA0040 KRT18 LRRC17 MATN3 MN1 MYB NAT1 NOVA1 NUP88 PIB5PA RARA
SLC1A1 SLC26A3 SLC39A6 SLC7A2 SMA3 SREBF1 TCEAL1 TFF1 TFF3 TGFB3 TUBA2 UGCG
ZBTB16
S48 S50 S53 S57 S65 S67 S71 S75 S12 S20 S24 S44 S80 S81 S83 S84 S85 S87 S88 S89 S90
S91 S92 S93 S96 S97 S98 S100
38 21
ABCG1 ACOX2 ALDH6A1 ARL3 ASAH1 BMPR1B CCNG2 CEACAM5 CRAT CRIP2 DOK1
DUSP4 DUSP5 GDF15 GLI3 GLRB GRCA HGD HMGCS2 HOXB2 HPX IGFBP4 LASP1 NBL1
NRIP1 OVGP1 PLCL1 PTPRN2 QDPR RAB31 RNASE4 RTN1 SCGB2A2 SEMA3C SERPINA5
SPARC TFAP2B UNG2
S48 S57 S65 S67 S12 S20 S24 S44 S80 S81 S83 S84 S87 S88 S89 S90 S91 S96 S97 S98
S100
6 18
BTG3 CENPA DSC2 LDHB SOD2 VGLL1
S48 S50 S53 S65 S75 S44 S80 S81 S83 S84 S85 S87 S88 S90 S96 S97 S98 S100
```

Figure 18: *bicluster results from plaid model it has total number of bicluster (**row 1**), title of the first Biclust result (**row 2**), dimension of the first biclust (row **3**), name of the first genes (**row 4-7**), And name of first the conditions (**row 8-9**), the same is for the remain two biclusters*

5 Discussion

In this project, we developed the first R package for diagnostic called BcDiag to; visualize profiles of single biclustered, biclusters vs. cluster rows or columns and to diagnostics the residuals of bicluster data

The R package can be downloaded from the CRAN repository (http://cran.r-project.org/web/packages/BcDiag/index.html). It has multi platform features and can be installed without any problem. We have applied different plots such as: 3D plot, Histograms, box plots and profile plots to visualize the data matrix. We have also included the residual plots obtained from two way ANOVA and median polish for bicluster data. It has also a feature of writing the bicluster results which are obtained from different algorithms.

The package has tested in two micro array datasets. It works well in both cases. This application can be used as alternative to visualize and diagnostic biclustered data in multi ways. Besides, the package supports different biclustered algorithms from three different packages. This will increase the efficiency of the exploratory data analysis. Moreover, it will save the time needed to write different codes for each algorithm. The application doesn't require advanced programming knowledge to use it. By changing the parameter values, it is possible to draw the required plot.

We used R software to develop the package for many reasons; it is open software, multi task, compatibility, multiple packages and built in functions and easy to understand. The package used traditional plots for visualization. However, we can have more features if we applied it in more advanced plots such as, ggplot2. The package can be extended by adding more features such as handling overlapped bicluster data and profiling more than one bicluster.

6 REFERENCES:

[1] Ben-Dor A., Chor B., Karp R., and Yakhini Z. (2003). Discovering local structure in gene expression data: The order-preserving submatrix problem proc. *Journal of Computational Biology*, **10**. 373–384.

[2] Madeira,S. and Oliveira,A. (2004). Biclustering algorithms for biological data. IEEE/ACM Trans. Comput. Biol. *Bioinform*., **1**, 24–45.

[3] Prelic, A. et al. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

[4] Vanmechelen, I. and Schepers, J. (2006): A unifying model for biclustering.
In: Compstat 2006 - *Proceedings in Computational Statistics*, 81–88.

[5] Sharan, R., Elkon, R. and Shamir, R. (2002). *Cluster analysis and its application to gene Expression data. Ernst Schering workshop on Bioinformatics and Genome Analysis*. Springer.

[6] Tukey, J.W. (1970). *Exploratory Data Analysis* (Limited Priliminary Edition), Volume II. Reading, MA: Addison-Wesley.

[7] Mood, A. M. (1950). *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

[8] Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

[9] Velleman, P.F. and Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*. Boston, MA: Duxbury Press.

[10] Murali, T. and Kasif, S. (2003): Extracting conserved gene expression motifs from gene expression. In*: Pacific Symposium on Biocomputing*, **8**,77–88.

[11] Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A. *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, **415**, 530-536.

[12] Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E. *et al*. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, *New Engl. J. Med*., **346**. 1937-1947.

[13] Hoshida, Y., Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2007). Subclass Mapping: Identifying Common Subtypes in Independent Disease Data Sets, *PLoS ONE*, **2**(11).

[14] Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and Zitzler, E. (2006). Bicat: a biclustering analysis toolbox. *Bioinformatics*, **22**, 1282–1283.

[15] Cheng, K.O., Law, N.F., Siu, W.C., and Lau, T.H. (2007). BiVisu: Software tool for bicluster detection and visualization. *Bioinformatics* , **23**, 2342-2344.

[16] Santamaria,R.,Theron R., and Quintales, L. (2008). BicOverlapper: A tool for bicluster visualisation. *Bioinformatics* , **24**, 1212-1213.

17] Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20.** 1993-2003.

[18] Ihmels J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat. Genet*, **31**, 370-377.

[19] Houglin, D.C., Mosteller, F., and Tukey, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons, Inc.

[20] Hochreiter, S., Bodenhofer, U., Heusel, M. *et al*. (2010). FABIA: factor analysis for bicluster acquisition. *Bioinformatices*, **26**, 1520-1527.

[21] Lazzeroni, L., Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica*, **12**, 61-86.

[22] Kaiser S. and Leisch F. (2008). A Toolbox for Bicluster Analysis in R. *Ludwigstrasse***. 33.**

[23] Tanay, A., Sharan, R ., and Shamir, R. (2002). Discovering Statistically Significant Biclusters in Gene Expression Data. *Bioinformatics* **18,** 196.205

7. Appendix

7.1 BcDiag Manual

Package 'BcDiag'

January 5, 2013

Version 1.0

Date 2012-11-15

Title Diagnostics plots for Bicluster Data

License GPL (>= 2)

Depends biclust, fabia, isa2

Imports methods

Author Aregay Mengsteab

Maintainer Aregay Mengsteab <mycs.zab@gmail.com>

Description This package provides Diagnostics plots for Bicluster data
obtained from packages; ''biclust'' by Kaiser et al.(2008),''isa2'' by Csardi et al. (2010)
and ''fabia'' by Hochreiter et al. (2010).

R topics documented:

Description

Bicluster Diagnostics plots

Introduction

The Bicluster Diagnostics plots(BcDiag) package is a visualization technique, for profiling and summarizing Bicluster data, particularly for gene expression level data. Target data matrix are bicluster genes(rows) and conditions(columns) versus clustered genes or conditions.

Main task

A BicDiag is a package of visualaization bicluster data, which is a subset matrix that have similar characterstics in terms of row(genes) and columns(conditions).

It has used three different types of bicluster algorithms to extract the biculsterd data; 'biclust','isa2' and 'fabia'. plots such as boxplot,histogram, line plot,3D plot are some of the plots that have used to visualize the data.

Major taskes of the package can be categorized in to three sections;

1. profiling and summarizing the biclustered vs. the clustered simultaneously
2. profiling and summarizing only the biclusterd data.
3. exploring the biclusterd data using anova and median polish techniques.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

References

Hochreiter, S., Bodenhofer, U., Heusel, M.et al. (2010).FABIA: factor analysis for bicluster acquisition. Bioinformatices, 26, 1520-1527.

Kaiser S. and Leisch F. (2008). A Toolbox for Bicluster Analysis in R. Ludwigstrasse. 33.

Csardi G., Kutalik Z., and Bergmann S.(2010). Modular analysis of gene expression data with R. Bioinformatics, 26, 1376-7

See Also

The Bicluster algorithms in the packages biclust,fabia and isa2.

anomedOnlybic The anomedOnlybic function

Description

Provides ANOVA and median polish residual plots for biclusterd data.

Usage

```
anomedOnlybic (dset,bres,fit="boxplot",mname="biclust",bnum=1)
```

Arguments

dset	Data matrix.
bres	Bicluster result.
fit	A string value to fit a plot; 'aplot','mplot','anovbplot','mpolishbplot','boxplot'
mname	Method name; 'biclust', 'isa2' or 'fabia'
bnum	Existed biclusters; '1','2'...

Details

A function provides residuals plots for biclustered data based on anova and median polish.

The function checked the required parameter values and fit the plot according to the user Requirement and generate an error message and suggestion if it is not fulfilled.

Value

residual plots or residual box plots.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

References

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer,Nature, 415, 530-536.

Kaiser S. and Leisch F. (2008). A Toolbox for Bicluster Analysis in R. Ludwigstrasse. 33.

See Also

plotOnlybic

Examples

```
data(breastc)
# find bicluster using one of biclust algorithms
bic<- biclust(breastc, method=BCPlaid())
#fit anova residual plot
anomedOnlybic(dset=breastc,bres=bic,fit="aplot",mname="biclust")
```

breastc Gene Expression Data Example

Description

Microarray data set of van't Veer breast cancer

Usage

```
data(breastc)
```

Format

A data matrix with 1213 genes and 97 samples.

References

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer,Nature, 415, 530-536.

Examples

```
data(breastc)

head(breastc)
```

dlbcl Gene Expression Data Example

Description

Log transformed Microarray data set of Rosenwald diffuse large-B-cell lymphoma

Usage

```
data(dlbcl)
```

Format

A data matrix with 661 genes and 141 samples.

References

Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, New Engl. J. Med., 346. 1937-1947.

Examples

```
data(dlbcl)

head(dlbcl)
```

---

exploreBic The exploreBic function

Description

Provides exploratory plots for biclusterd and. clusterd data parallely.

Usage

```
exploreBic(dset, bres, gby ="genes", pfor ="mean", mname ="biclust", bnum =1)
```

Arguments

dset	Data matrix.
bres	Bicluster result.
gby	group bicluster; 'genes' or 'conditions'.
pfor	plot for 'mean', 'median', 'variance', 'mad', 'all', or 'quantile'.
mname	Method name; 'biclust', 'isa2' or 'fabia'
bnum	Existed biclusters; '1','2'...

Details

The exploreBic function is mainly used for exploratory data analysis. It provides summary plots for 'mean' 'median', 'variance', 'mad' and 'quantile plot'.

The exploreBicfunction checked if the parameters are appropriately submitted and then it calls the indexedBicfunction to identify the biclust sub matrix and forwarded it to exploreCalc to be calculated its summary statistics. Finally, explorePlot will be invoked to display the required plot.

Value

Summary plot will display according to the user specification.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

References

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer,Nature, 415, 530-536.

Hochreiter, S., Bodenhofer, U., Heusel, M.et al. (2010).FABIA: factor analysis for bicluster acquisition. Bioinformatices, 26, 1520-1527.

See Also

exploreOnlybic

Examples

```
data(breastc)
# find bicluster using fabia algorithm
fab<- fabia(breastc)
# Plot the mean of biclusterd and clustered genes parallely.
exploreBic(dset=breastc,bres=fab,gby="conditions",pfor="mean",mname="fabia")
```

---

exploreCalc The exploreCalc function

Description

calculates the 'mean', 'median','variance', 'MAD', and 'quantile' for biclusterd and. clusterd data.

Usage

```
exploreCalc(bic.mat)
```

Arguments

bic.mat	Biclusterd or clustered data matrix.

Details

an internal function used to calculate the summary statistics of biclustered and clustered data. it supports for exploreBic function. It is used only for internal purpose.

Value

Summary statistics('mean', 'median','variance', 'MAD', and 'quantile') for biclusterd and clusterd data.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

See Also

exploreBic

exploreOnlybic The exploreOnlybic function

Description

Provides exploratory plots only for biclusterd.

Usage

```
exploreOnlybic(dset, bres, fit= "all", gby= "genes", mname="biclust",bnum=1)
```

Arguments

dset	Data matrix.
bres	Bicluster result.
gby	group bicluster; 'genes' or 'conditions'.
fit	fit a plot for 'mean', 'median', 'variance', 'mad', 'all', or 'quantile'.
mname	Method name; 'biclust', 'isa2' or 'fabia'
bnum	Existed biclusters; '1','2'...

Details

The exploreOnlybic function has similar function with exploreBic. the only difference is, it provides exploratory plots only for biclusterd data. It provides summary plots for 'mean' 'median', 'variance', 'mad' and 'quantile plot'.

Value

Summary plot will display only for biclusterd data.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

References

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer,Nature, 415, 530-536.

Hochreiter, S., Bodenhofer, U., Heusel, M.et al. (2010).FABIA: factor analysis for bicluster acquisition. Bioinformatices, 26, 1520-1527.

See Also

exploreBic

Examples

```
data(breastc)
# find bicluster using fabia algorithm
fab<- fabia(breastc)
# Plot the median of biclusterd data.
exploreOnlybic(dset=breastc,bres=fab,fit="median",gby="genes",mname="fabia",bnum=1)
```

explorePlot The explorePlot function

Description

An internal function supports for exploreBic and Provides different types of plots for biclustered and clustered data simultaneously.

Usage

```
explorePlot(sbic, obic, pfor = c("all", "mean", "median",
"variance", "mad", "quant"))
```

Arguments

sbic	biclusterd data
obic	clusterd data
pfor	plot for 'mean', 'median', 'variance', 'mad', 'all', or 'quantile'.

Details

It has an important role to plot the summary statistics of biclustered and clustered data. it helps the exploreBic function by exracting different types of plots for the summary statistics of biclusterd and clusterd data.

Value

Extracts the required summary plot.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

See Also

exploreBic,exploreCalc

indexedBic The indexedBic function.

Description

Supportive function for profileBic, exploreBic, exploreOnlybic and anomedOnlybic used to extract biclusterd index from a dataset based on 'biclust','fabia' and 'isa2' algorithms.

Usage

```
indexedBic(dset, bres, mname = c("fabia", "isa2", "biclust"), bnum)
```

Arguments

dset	Data matrix
bres	Bicluster result
mname	Method name; 'biclust', 'isa2' or 'fabia'
bnum	Existed biclusters; '1','2'...

Details

Its main purpose is for internal usage and extracts the index of biclusterd rows and columns from the dataset and returns to the main function. by knowing the indexed of the biclusterd data, we can easily identified and grouped the biculsterd data from the clusterd ones.

Value

Indicies of biclustered data.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

See Also

profileBic, exploreBic, exploreOnlybic, anomedOnlybic

---

plotOnlybic The plotOnlybic function

Description

Supportive function for exploreBic used to plot summary statistics of biclusterd data.

Usage

```
plotOnlybic(ball, fit = "all", gby)
```

Arguments

ball	List of summary statistics('mean','median','variance','mad').
fit	fit a plot for 'mean', 'median', 'variance', 'mad', 'all', or 'quantile'.
gby	grouped by 'genes' or 'conditions'.

Details

A supportive function provides plots for the summary statistics of biclusterd data such as 'mean','median','variance', 'mad' or 'all'. if 'all' is selected then it will be displayed all the summary statistics together in single frame.

Value

Summary plot according to the user requirement

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

See Also

exploreOnlybic

---

profileAll The profileAll function

Description

Supportive function for profile.bic used to plot all profile plots in single frame for biclusterd and clusterd data.

Usage

profileAll(dset, indg, indc, grp, gby = "genes", teta = 120, ph = 30)

Arguments

dset	Data Matrix
indg	indexed genes(rows) obtained from indexedBic
indc	indexed conditions(columns) obtained from indexedBic
grp	group genes(rows)/conditions(columns) as ’1’ for biclusterd or ’2’ out side biclusterd
gby	grouped by ’genes’ or ’conditions’
teta	numerical value to rotate the 3D; 0,90,180,..
ph	numerical value to rotate the 3D; 0,90,180,..

Details

profileAll is a function which supports profileBic and provides all profile plots; ’line’,’histogram’, ’boxplot’,’3D’ of the biclusterd and clusterd data grouped either by ’genes’ or ’conditions’ and displayed in a single frame.

Value

Profile plots for all biclusterd and clustered genes and conditions.

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

See Also

profileBic

profileBic	The profileBic function.

Description

Provides profile plots for biclusterd and clusterd data parallely.

Usage

```
profileBic(dset, bres, mname = c("fabia", "isa2", "biclust"), bplot = "all",
gby = "genes", bnum = 1, teta = 120, ph = 30)
```

Arguments

dset	Data matrix.
bres	Bicluster result.
mname	Method name; 'biclust', 'isa2' or 'fabia'
bplot	types of plots;'all','lines','boxplot','histogram' or '3D'
gby	grouped by; 'genes', or 'conditions'
bnum	Existed biclusters; '1','2'...
teta	numerical value to rotate the 3D; 0,90,180,..
ph	numerical value to rotate the 3D; 0,90,180

Details

The profile.bic function checks all parameters are correctly submitted then it calls a function indexedBic to identify the biclusterd and clusterd data.

Value

profile.bic (dset, bres, mname="biclust", bplot="all", gby="genes", bnum=1, teta=120, ph=30)

Author(s)

Mengsteab Aregay <mycs.zab@gmail.com>

References

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer,Nature, 415, 530-536.

Kaiser S. and Leisch F. (2008). A Toolbox for Bicluster Analysis in R. Ludwigstrasse. 33.

See Also

profileAll

Examples

```
#find biclust result
data(breastc)
bic<- biclust(breastc, method=BCPlaid())# 3 bicluster found.
# 3D profile plot for biclusterd and clusterd data
# grouped by conditions.
profileBic(dset=breastc,bres=bic,mname="biclust",
bplot="3D",gby="conditions",teta=-30,ph=50,bnum=1)
```

---

writeBic The writeBic function

Description

Provides a summary output in a text format, extracted from ’biclust’,’isa2’ and ’fabia’ bicluster algorithms.

Usage

```
writeBic(dset, fileName, bicResult, bicname,
mname = c("fabia", "isa2", "biclust"), append = TRUE, delimiter = " ")
```

Arguments

dset	Data matrix
fileName	The name of the bicluster file to be saved.
bicResult	bicluster result obtained from ’biclust’,’isa2’ or ’fabia’
bicname	the tilte to be given for the biclusterd data.
mname	method name; ’biclust’, ’isa2’ or ’fabia’
append	logical value; ’true’ or ’false
delimiter	default value is " ".

Details

The original function was developed in ’biclust’ package by Kaiser et.al (2008). we extend the function to be used for further bicluster algorithms, such as; ’isa2’ and ’fabia’.

Value

Biclusterd text file with title, total number of biclusterd, dimention and name of the biclusterd genes(rows) or conditions(columns).

Author(s)

Mengsteab Aregay
<mycs.zab@gmail.com>

References

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer,Nature, 415, 530-536.

Kaiser S. and Leisch F. (2008). A Toolbox for Bicluster Analysis in R. Ludwigstrasse. 33.

Csardi G., Kutalik Z., and Bergmann S.(2010). Modular analysis of gene expression data with R. Bioinformatics, 26, 1376-7

See Also

biclust

Examples

```
#manupilate the biclust result
data(breastc)
isa<- isa(breastc)
#write the bicluster result in to a text
writeBic(dset=breastc,fileName="isabreast.txt",
bicResult=isa, bicname="Biclust results for isa",
mname="isa2")
```

7.2 BcDiag algorithms for all functions

ProfileBic algorithm:

1 ***Check:*** *all parameters are submitted correctly.*
2 ***Call:*** *the function* ***'indexedBic'*** *and identify bicluster from the data set using the indices*
3 ***Group:*** *the data matrix according to user specification.*
4 ***Sort:*** *the biclust in one group and the non biclust in other group.*
5 ***Check:*** *the parameter '****bplot'*** *and if* ***'all'*** *is required, call the function* ***'profileAll'.***
6 ***Plot:*** *the required output plot for visualization.*

ExploreBic algorithm:

1 ***Check:*** *if all parameters are submitted correctly.*
2 ***Call:*** *the function* ***'indexedBic'*** *and identify bicluster from the data set using the indices*
3 ***Group:*** *the data matrix according to user specification.*
4 ***Sort:*** *the biclust in one group and the non biclust in other group.*
5 ***Call:*** *the function* ***'exploreCalc'*** *and calculate the '****mean', 'variance'****, '****median', 'quantile'*** *and* ***' mad'.***
6 ***Plot:*** *the required graph by calling the function* ***'explorePlot'.***

ExploreOnlybic algorithm:

1 ***Check:*** *all parameters are submitted correctly.*
2 ***Call:*** *the function* ***'indexedBic'*** *and identify bicluster from the data set using the indices*
3 ***Group:*** *the data matrix according to user specification (either row wise or column wise Bicluster).*
4 ***Repeat:*** *calculate the summary statistics either row wise or column wise, iteratively.*
5 ***Call:*** *the function'* ***plotOnlybic'*** to draw the plot.

AnomedOnlybic algorithm:

1. ***Check:*** *all parameters are submitted correctly.*
2. ***Call:*** *the function* ***'indexedBic'*** *and identify bicluster from the data set using the indices*
3. ***fit:*** *ANOVA model and median polish and obtain the results*
4. ***Plot:*** *the residuals according to the user interest.*

WriteBic algorithm:

1. ***Check:*** *all parameters are submitted correctly.*
2. ***Check:*** *the name of the method*
3. ***Repeat:*** *number of the biclust times to write all biclusters.*
4. ***Write:*** *the biclustered in a text format.*

7.3. Breast Cancer Dataset

i. Biclust plaid model

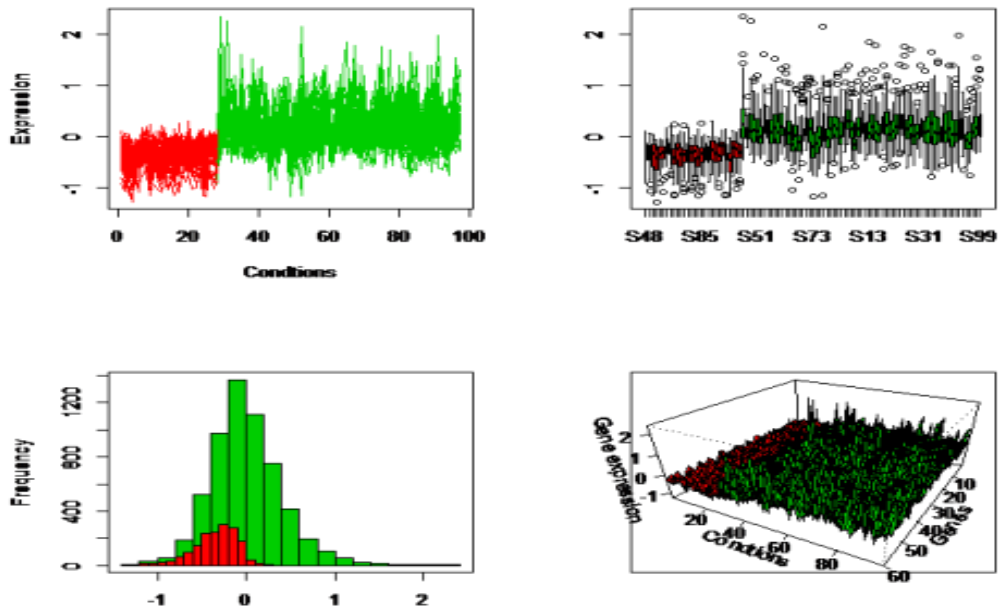


Figure a.1: *Profile plot for the first biclustered gene expression data based on plaid Model algorithm, grouped by conditions*

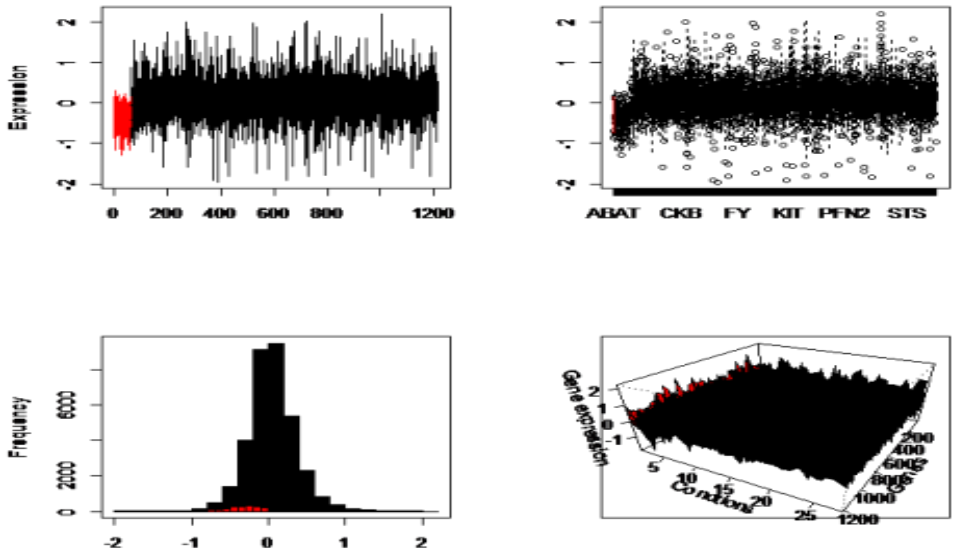


Figure a.2: *Profile plot for the first biclustered gene expression data based on plaid Model algorithm, grouped by genes*

ii. ISA algorithm

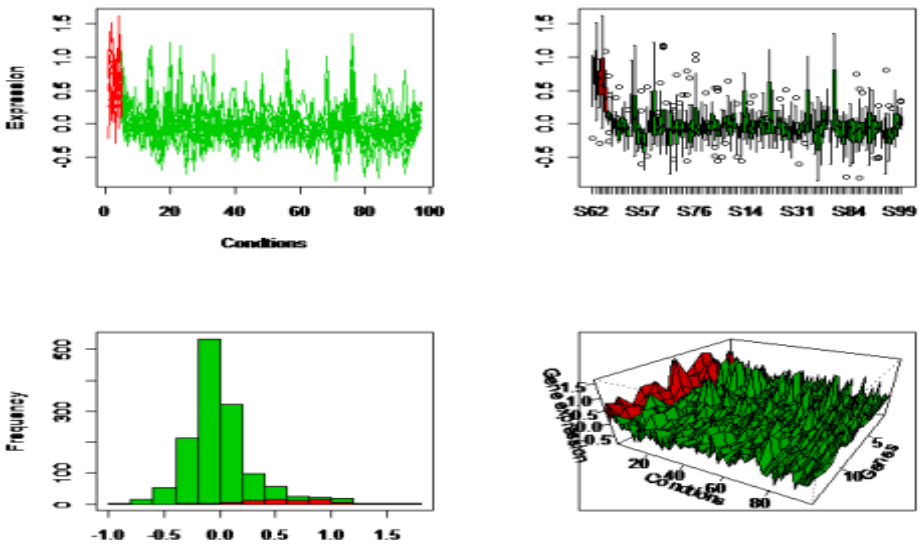


Figure b.1: *Profile plot for the first biclustered gene expression data based ISA (Iterative search algorithm) algorithm, grouped by conditions*

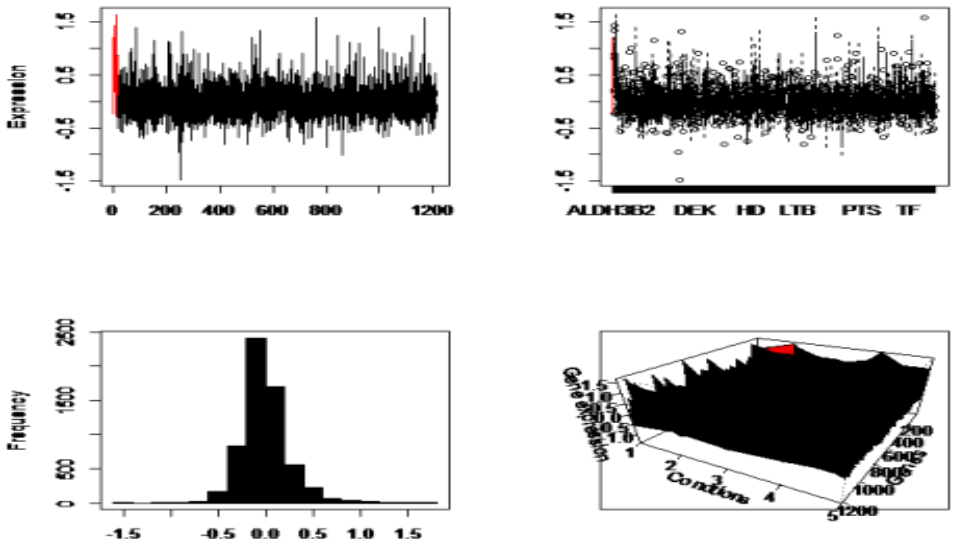


Figure b.2: *Profile plot for the first biclustered gene expression data based ISA Algorithm, grouped by genes*

b. Exploratory biclustered genes using summary statistics.

i. Biclust plaid model algorithm for exploratory plots

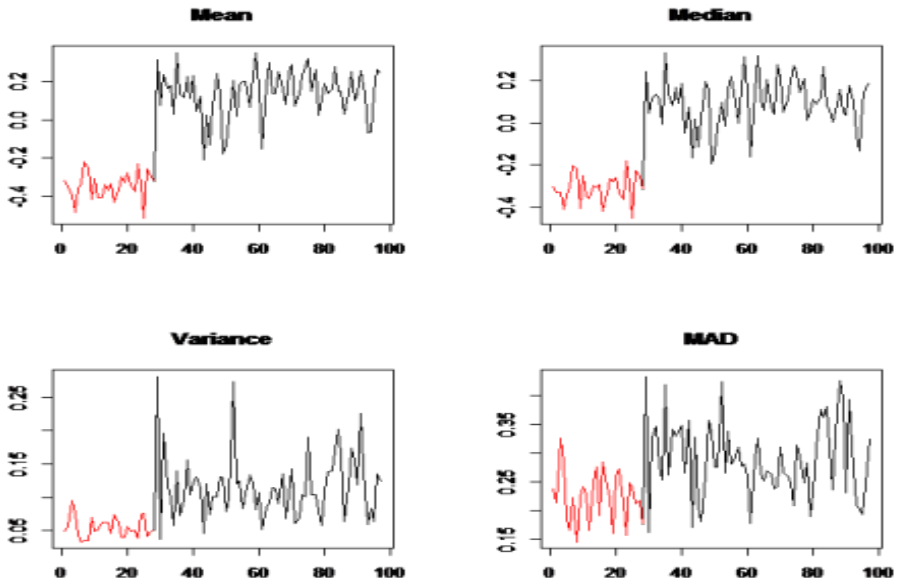


Figure c.1: *Summary statistics biclustered gene data for an output of plaid model Algorithm grouped by conditions*

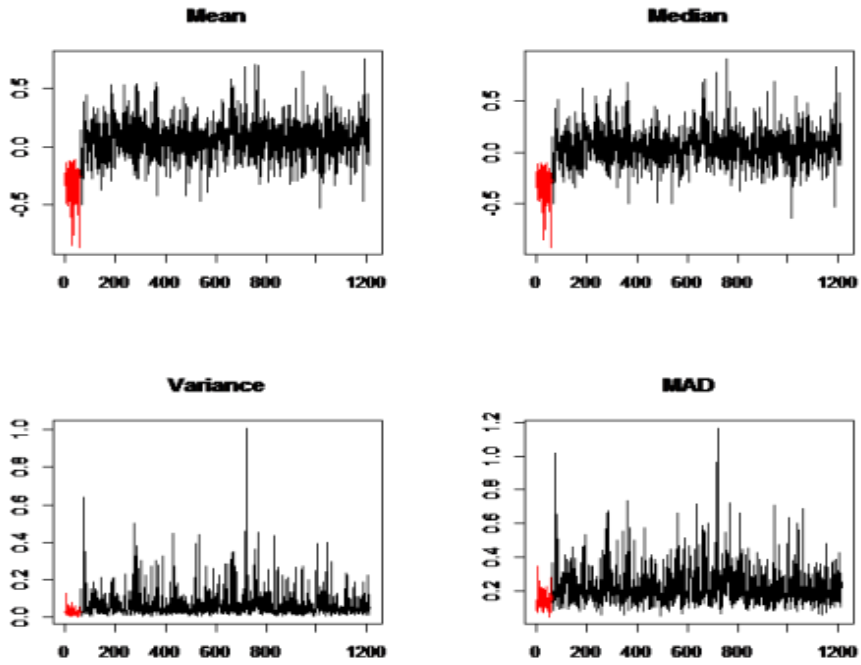


Figure c.2: *Summary statistics biclustered gene data for an output of plaid model Algorithm grouped by genes*

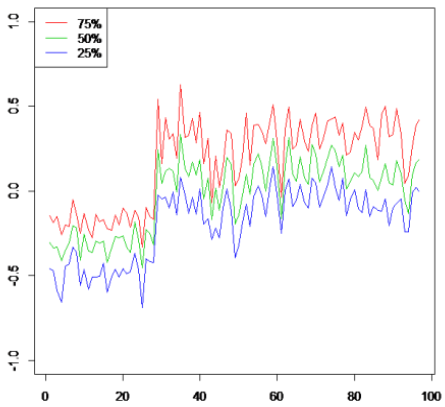


Figure d.1: *Quantile plot grouped by conditions using plaid model algorithm*

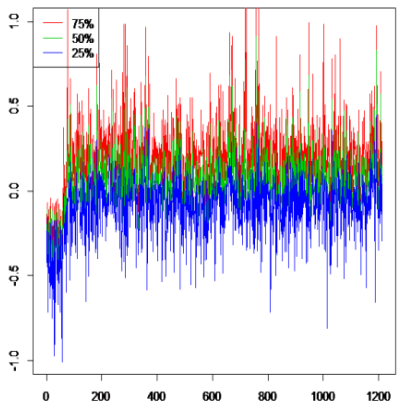


Figure d.2: *Quantile plot grouped by genes using plaid model algorithm*

ii. ISA algorithm for exploratory plots.

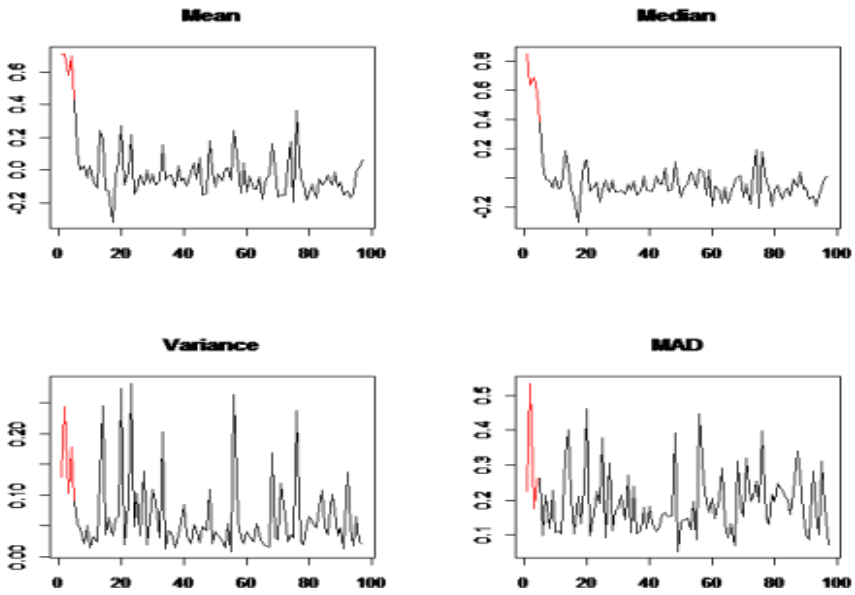


Figure e.1: *Summary statistics biclustered gene data for an output ISA algorithm grouped By Conditions*

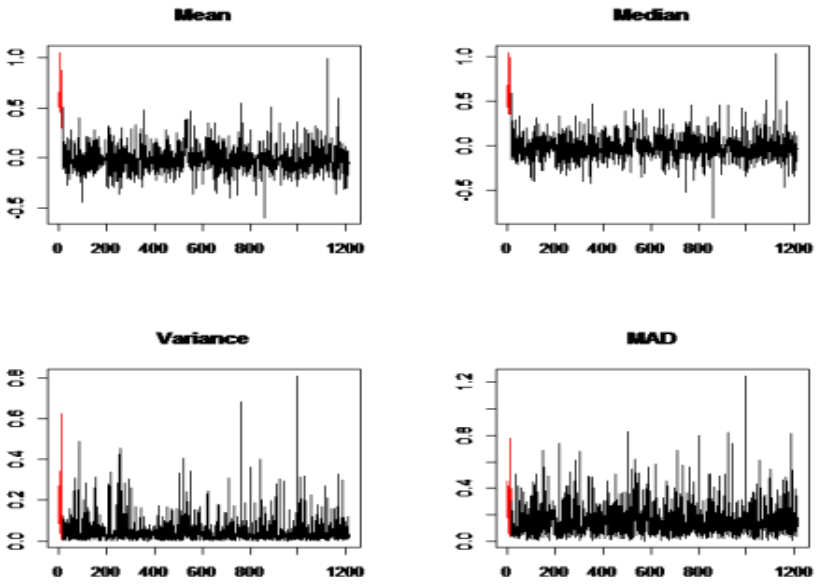


Figure e.2: *Summary statistics biclustered gene data for an output of ISA algorithm Grouped by genes*

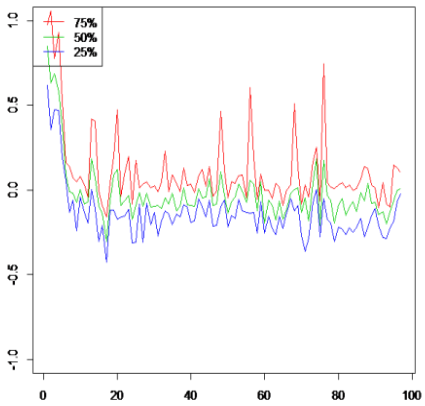


Figure f.1: *Quantile plot grouped by conditions using ISA algorithm*

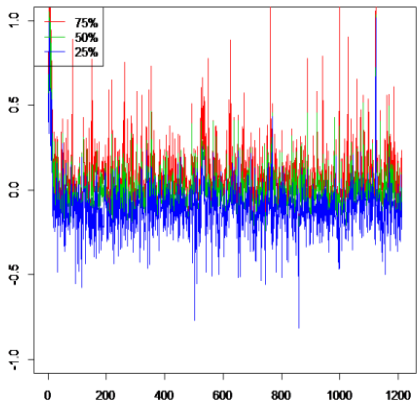


Figure f.2: *Quantile plot grouped by genes using ISA algorithm*

c. Exploratory plot only for biclustered genes and conditions

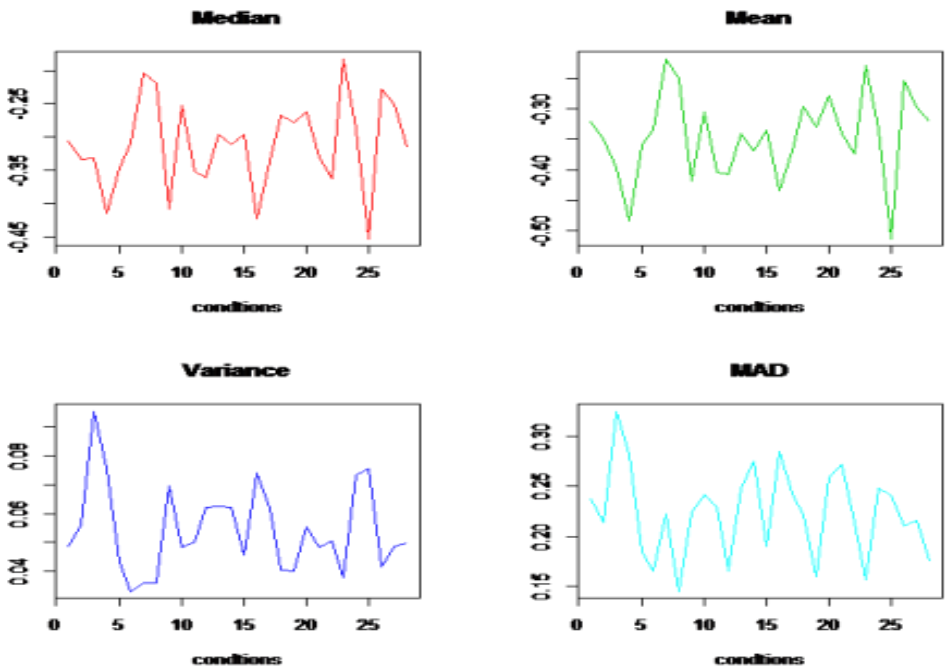


Figure g.1: *Exploratory plot only for biclustered conditions from biclust plaid model algorithm*

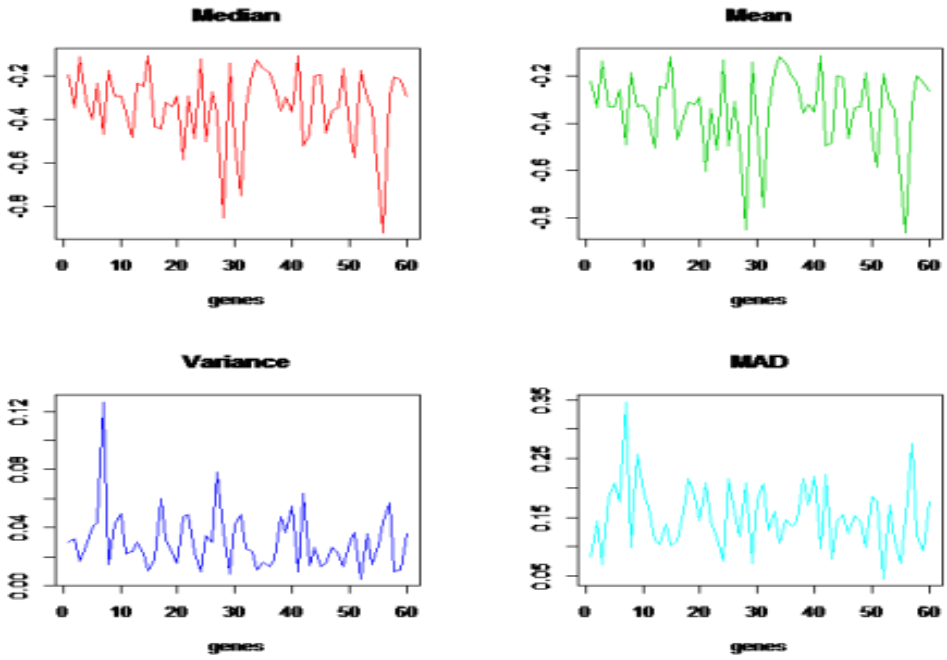


Figure g.2: *Exploratory plot only for biclustered genes using the biclust plaid model Algorithm*

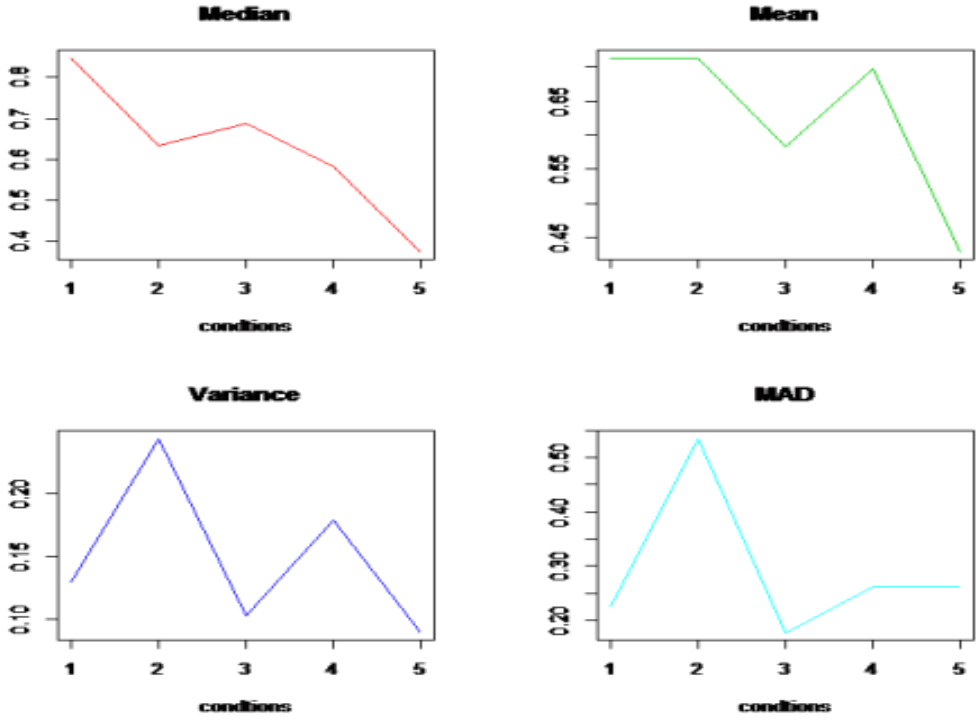


Figure h.1: Exploratory plot only for biclustered conditions from ISA algorithm

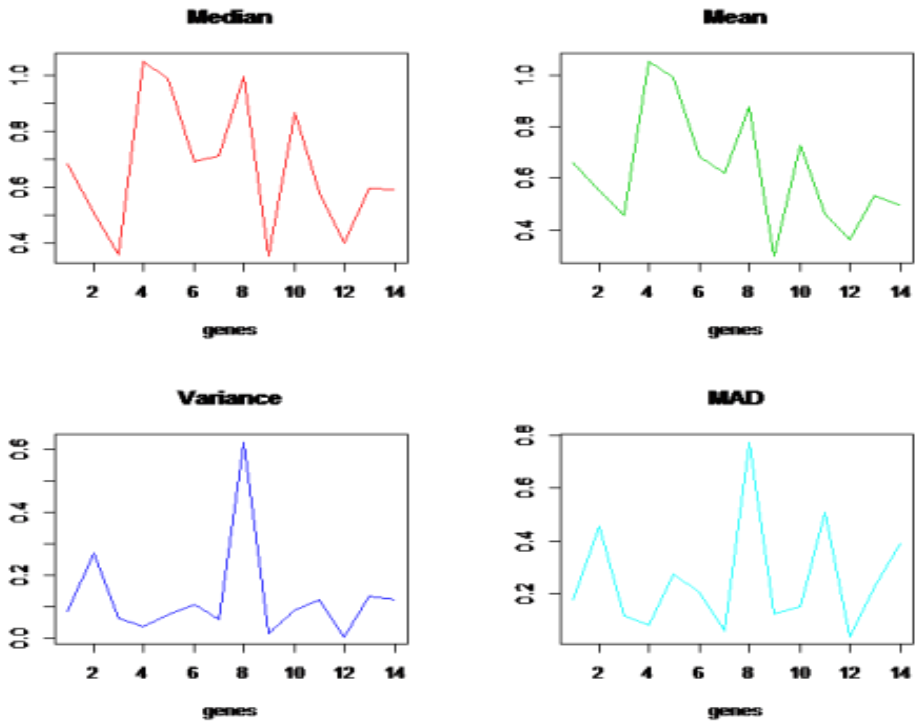


Figure h.2: *Exploratory plot only for biclustered genes from ISA Algorithm* Anomed Function.

i. Residual plots for biclust plaid model algorithms.

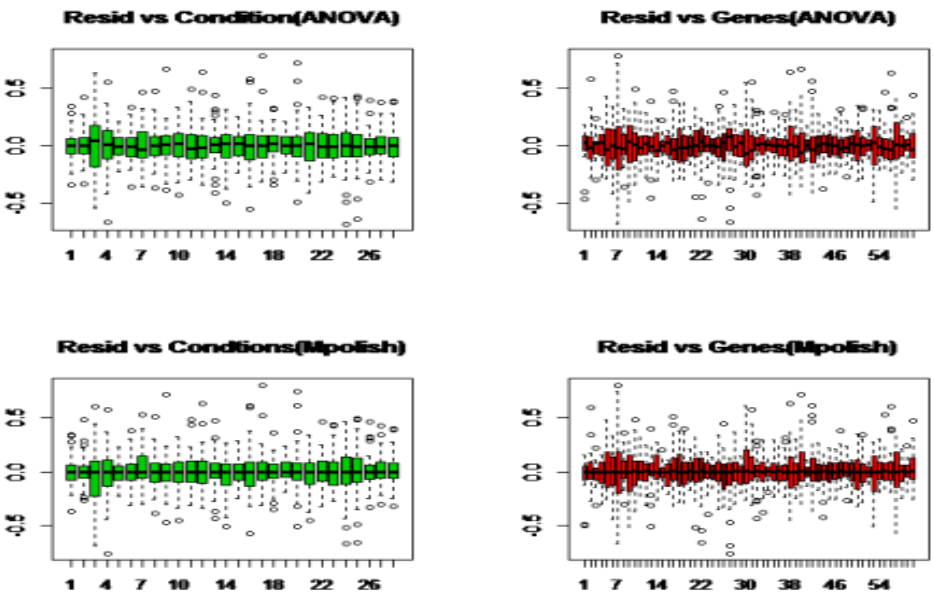


Figure i.1: *Box plots for plaid biclust results using ANOVA* (***top two***) *and median polish* (***Bottom two***) *for residuals vs. genes and condition*

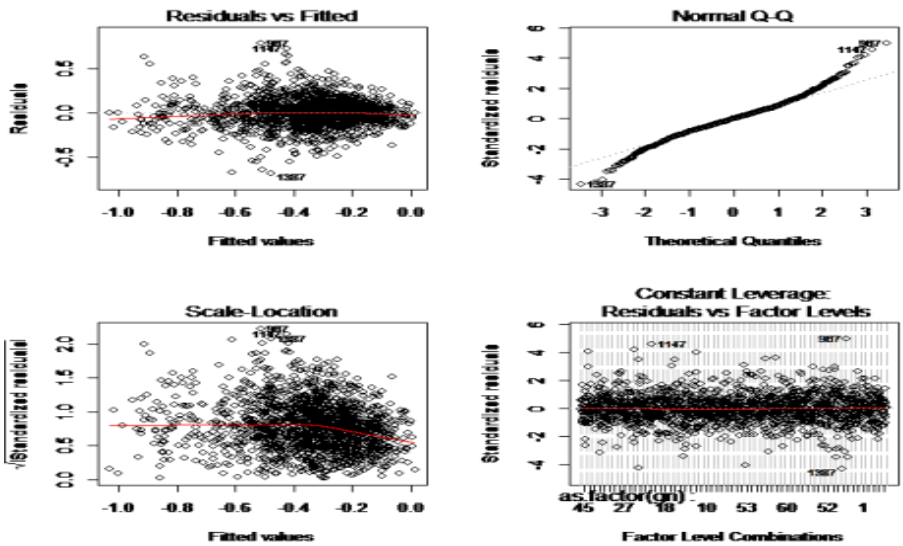


Figure i.2: *Residual plots for plaid model biclust results using ANOVA (**top two**) to check the Assumptions for constant variance (**top left**), normality (**top right**), outliers against y axis (**Bottom left**) and outliers against x axis (**bottom right**)*

ii. Residual plots for ISA algorithm

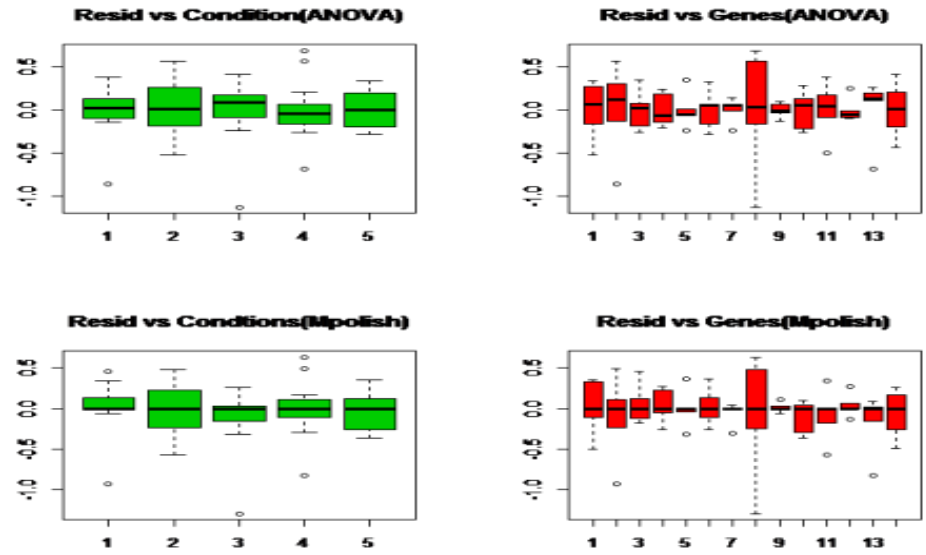


Figure j.1: *Box plots for ISA biclust results using ANOVA (**top two**) and median polish (**Bottom two**) for residuals vs. genes and conditions*

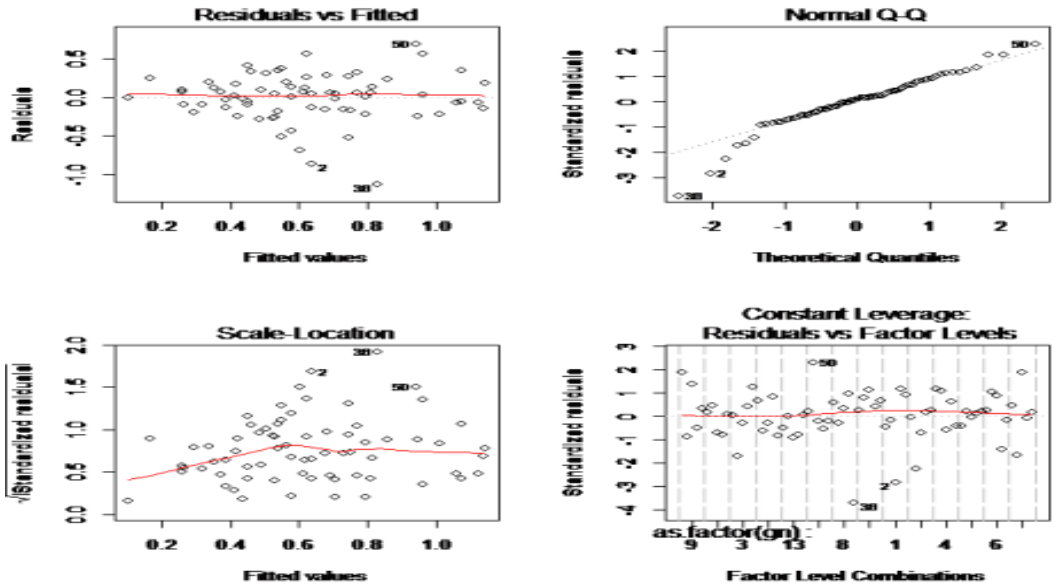


Figure i.2: *Residual plots for ISA biclust results using ANOVA (**top two**) to check the assumptions for Constant variance (**top left**), normality **(top right)**, outliers against y axis (**bottom left**) and outliers against x Axis (**bottom right**)*

7.4. DLBCL data

d. Profile plot for ISA algorithms

i. Grouped by conditions

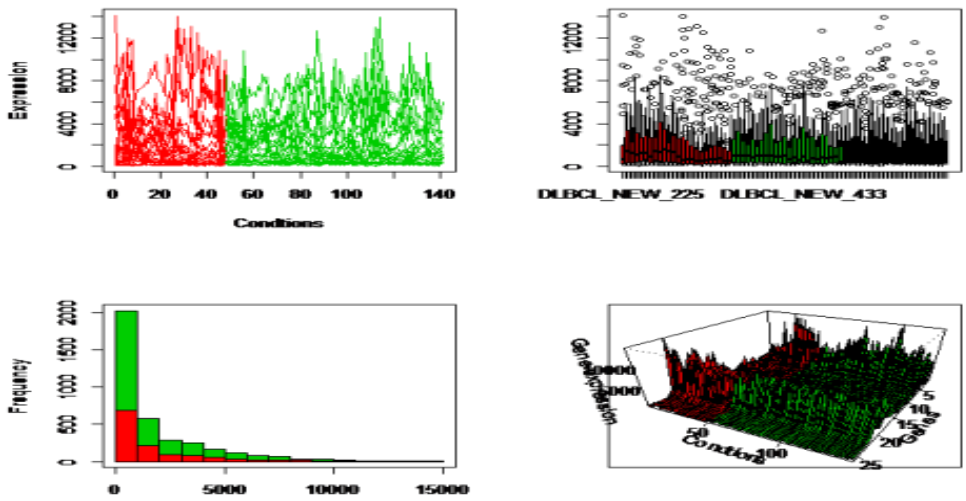


Figure j.1: *Profile plot for the first biclustered gene expression data based ISA algorithm, grouped by conditions.*

ii. Grouped by genes

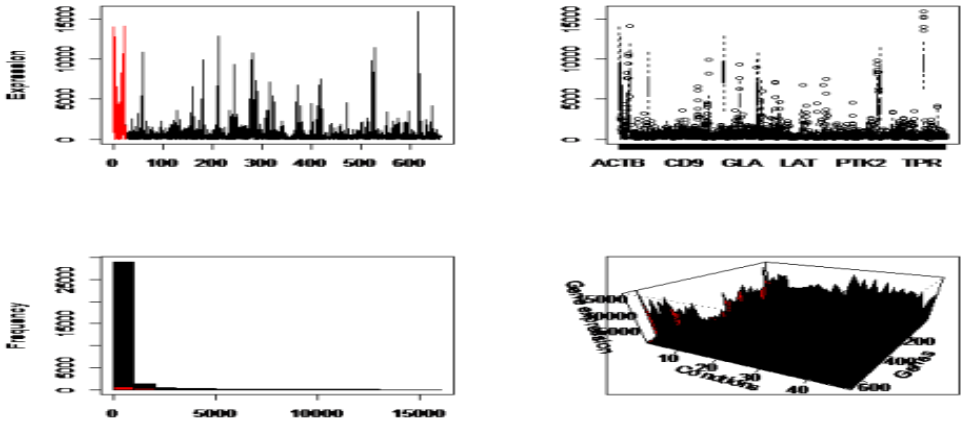


Figure j.2: *Profile plot for the first biclustered gene expression data based on ISA, grouped by genes.*

e. Exploratory plots for biclustered data.

i. Group by conditions

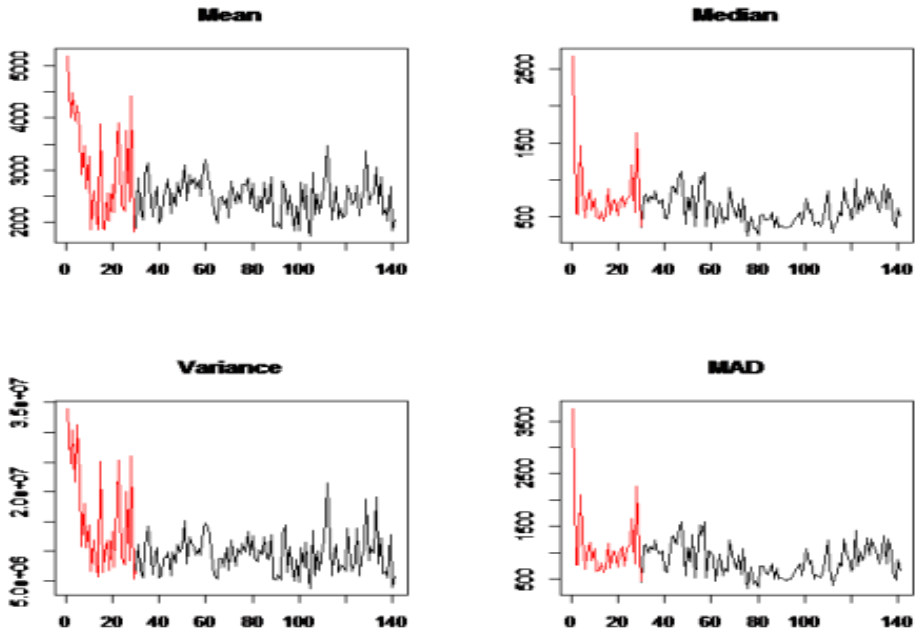


Figure k.1: *Summary statistics biclustered gene data for an output ISA algorithm grouped by conditions*

ii. Group by genes.

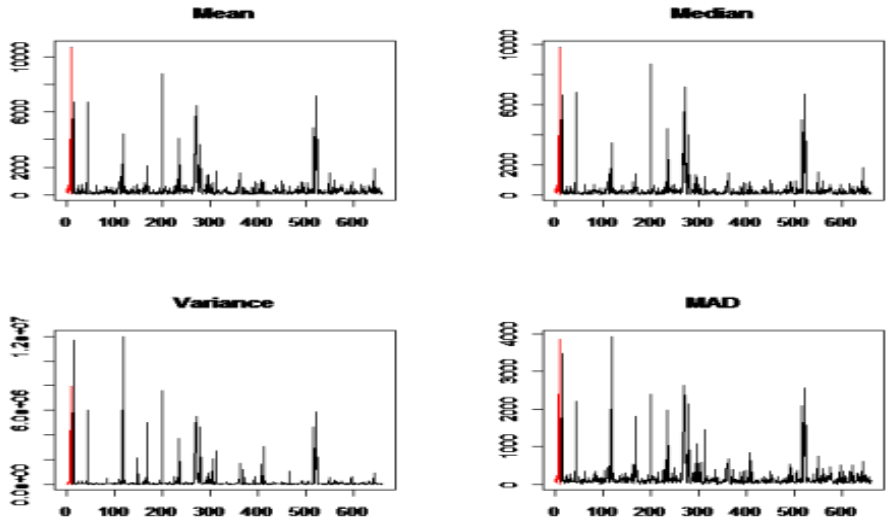


Figure k.2: *Summary statistics biclustered conditions data for an output of ISA algorithm grouped by genes*

f. Diagnostics plot only for biclustered data.

i. *For conditions*

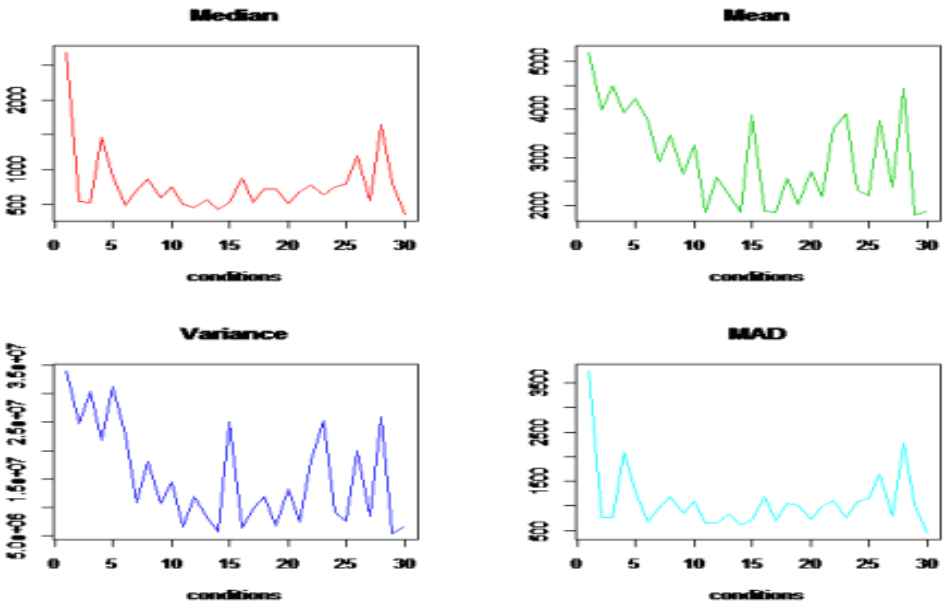


Figure l.1: *Exploratory plot only for biclustered conditions from ISA algorithm*

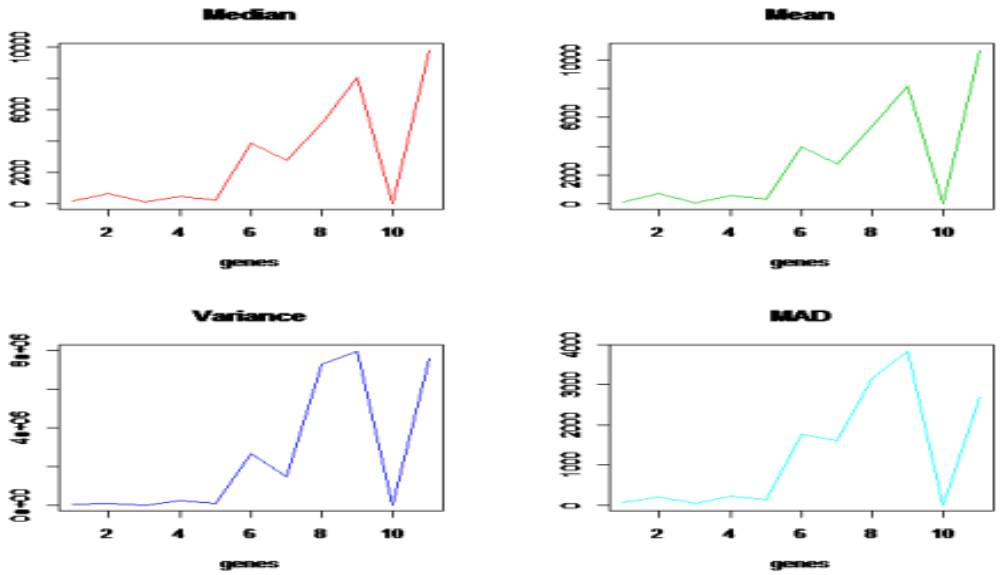


Figure l.1*: Exploratory plot only for biclustered genes from ISA algorithm*

2. Residual plots

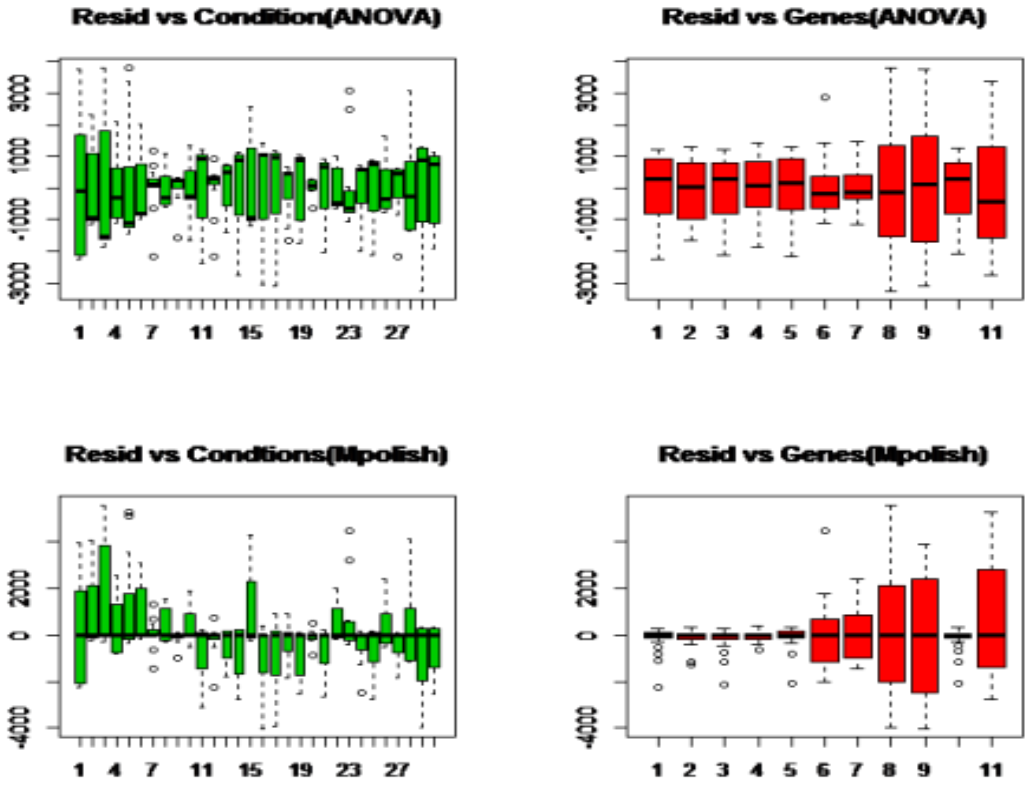


Figure m: *Box plots for ISA results using ANOVA (**top two**) and median polish (**Bottom two**) for residuals vs. genes and conditions*

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Diagnostic tools for biclustering output

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Aregay, Mengsteab Fantahu

Datum: **4/02/2013**