Journal of
Data Mining in Genomics & Proteomics

Yadav and Srivastava, J Data Mining Genomics Proteomics 2016, 7:3
DOI: 10.4172/2153-0602.1000197

Review Article Open Access

Visualization of High Throughput Genomic Data Using R and Bioconductor

Ruchi Yadav and Prachi Srivastava*

Amity Institute of Biotechnology, Amity University, Uttar Pradesh, Lucknow, India

Abstract

DNA microarrays, technology aims at the measurement of mRNA levels in particular cells or tissues for many genes simultaneously. Microarray in molecular biology results in huge datasets that need rigorous computational analysis to extract biological information that lead to some conclusion. From printing of microarray chip to hybridization and scanning process it results in variability in quality of data due to which actual information is either lost or it is over represented. Computational analysis plays an important part related to the processing of the biological information embedded in microarray results and for comparing gene expression result obtained from different samples in different condition for biological interpretation. A basic, yet challenging task is quality control and visualization of microarray gene expression data. One of the most popular platforms for microarray analysis is Bioconductor, an open source and open development software project for the analysis and comprehension of genomic data, based on the R programming language. This paper describes specific procedures for conducting quality assessment of Affymetrix Gene chip using data from GEO database GSE53890 and describes quality control packages of bioconductor with reference to visualization plots for detailed analysis. This paper can be helpful for any researcher working on microarray analysis for quality control analysis of affymetrix chip along with scientific interpretations.

Keywords: Microarray; R; Bioconductor; Transcriptomics; Quality control; Genome visualization

Introduction

In the context of the human genome project, new technologies emerged that facilitate the parallel execution of experiments on a large number of genes simultaneously. The measurement of transcriptional activity in living cells is of fundamental importance in many fields of research from basic biology to the study of complex diseases such as cancer [1]. The so-called DNA microarrays, or DNA chips, constitute a prominent example. This technology aims at the measurement of mRNA levels in particular cells or tissues for many genes at once [2]. DNA microarrays provide an instrument for measuring the mRNA abundance of tens of thousands of genes. Currently, the measurements are based on mRNA from samples of hundreds to millions of cells, thus expression estimates provide an ensemble average of a possibly heterogeneous population [3].

Gene expression profiling provides unprecedented opportunities to study patterns of gene expression regulation, for example, in diseases or developmental processes. DNA microarray technology takes advantage of hybridization properties of nucleic acid and uses complementary molecules attached to a solid surface, referred to as probes single strands of complementary DNA for the genes of interest-which can be many thousands are immobilized on spots arranged in a grid (array) on a support which will typically be a glass slide, a quartz wafer, or a nylon membrane. From a sample of interest, e.g. a tumor biopsy, the mRNA is extracted, labeled and hybridized to the array [4]. Measuring the quantity of label on each spot then yields an intensity value that should be correlated to the abundance of the corresponding RNA transcript in the sample [5]. Microarrays provide a rich source of data on the molecular working of cells. Each microarray reports on the abundance of tens of thousands of mRNAs. Virtually every human disease is being studied using microarrays with the hope of finding the molecular mechanisms of disease [6-8].

Bioinformatics analysis plays an important part of processing the information embedded in large-scale expression profiling studies and for laying the foundation for biological interpretation [8-10].

A basic, yet challenging task in the analysis of microarray gene expression data is the identification of changes in gene expression that are associated with particular biological conditions. Careful statistical design and analysis are essential to improve the efficiency and reliability of microarray experiments throughout the data acquisition and analysis process [11-12].

Microarray Studies

Microarrays are useful in a wide variety of studies with a wide variety of objectives. Many of these objectives fall into the following categories [13,14].

a) A typical microarray experiment is one who looks for genes deferentially expressed between two or more conditions. That is, genes which behave differently in one condition (for instance healthy [or untreated or wild type] cells) than in another (for instance tumor [or treated or mutant] cells). These are known as class comparison experiments.

b) When the emphasis is on developing a statistical model that can predict to which class a new individual belongs we have a class prediction problem. Examples of this are predicting the response to a treatment (e.g. classes are _responder_ and _non-responder_) or the evolution of a disease (e.g. recidivated or cured).

c) Sometimes the objective is the identification of novel sub-types of individuals within a population. For example it has been shown

***Corresponding author:** Prachi Srivastava, Assistant Professor, Department of Biotechnology, Amity University, Uttar Pradesh, Lucknow, India, Tel: +919453141916; E-mail: psrivastava@amity.edu; p.srivastava2@gmail.com

Received May 02, 2016; **Accepted** May 17, 2016; **Published** May 26, 2016

Citation: Yadav R, Srivastava P (2016) Visualization of High Throughput Genomic Data Using R and Bioconductor. J Data Mining Genomics Proteomics 7: 197. doi:10.4172/2153-0602.1000197

that certain types of leukemia present some subclasses that are very hard to distinguish morphologically but which can be classified using gene expression. This is an example of class discovery.

d) Pathway Analysis studies are those that try to find genes whose co regulation reflects their participation in common or related biochemical processes.

One of the most popular platforms for microarray analysis is Bioconductor, an open source and open development software project for the analysis and comprehension of genomic data, based on the R programming language. This paper describes specific procedures for conducting quality assessment of Affymetrix Gene Chip using data from GEO database using the open-source R programming environment in conjunction with the open-source Bioconductor software [15].

Bioconductor and R

R is a programming language. The name "R" is initials of names of the two R authors (Robert Gentleman and Ross Ihaka). R is introduced in 1991 and R 1.0.0 is released in year 2000. Bioconductor emerges as a boon in life sciences and in high throughput experiments where analysis tools are available free of cost to analyze experimental data. In year 2008 Bioconductor version 2.4 is released and further follows R release. Current version of the Bioconductor is 3.2 and R version is 3.2.2 [16].

R environment is easy to use, coherent and have tools for data analysis. What make R different from other programming languages is its GUI for quick and easy upload of data along with tools for data manipulation, calculation and analysis along with it statistical tools facilitates calculation of standard deviation, variance, t-test, f-test and other statistical tools. R can be accessed from http://cran.r-project.org/ [17].

Bioconductor (www.bioconductor.org) provides bioinformatics tool for analyzing high throughput data that comes from experiment like microarray, SAGE, MS, MS-MS. Bioconductor data packages are divided into three categories Annotation Data, Experiment Data and Software. Currently there are 1104 software packages, 898 Annotation Data and 257 Experiment Data. It is very hard to identify particular package for set of experiments. This paper reviews the methods for visualization of affymetrix gene expression data. These steps are essential part of microarray data analysis that should be taken before utilized in processing and analysis of gene expression differential expression analysis [18].

Microarray data produces lots of experimental errors that emerged because of biasness in dye intensities, laser scanner, spotting errors, hybridization biasness. Before microarray analysis data must be cleaned and processed to extract biological information. Visualization and graphs representation are best suited to study microarray intensity files and comparing probe hybridization signals [19].

Evaluation of data quality is a major issue in microarray analysis. There are many packages that can be used for quality control analysis Table 1 lists the bioconductor packages that are used to study quality of chips and visualizing high throughput microarray data [20].

After quality control analysis and normalization differentially expressed genes calculation can be done for biological interpretation. Table 2 list the differentially expressed gene packages available at bioconductor [21].

Materials and Methods

The raw data for this study is retrieved from Gene Expression Omnibus database: http://www.ncbi.nlm.nih.gov/geo/; GEOID: GSE53890; http://

S. No	Package	Description
1	a4	Automated Affymetrix Array Analysis
2	a4Base	analysis of Affymetrix microarray experiments
3	a4Core	Automated Affymetrix Array Analysis
4	a4Preproc	package for preprocessing of microarray data
5	a4Reporting	Automated Affymetrix Array Analysis Reporting Package
6	affxparser	Ackage for parsing Affymetrix files (CDF, CEL, CHP, BPMAP, BAR)
7	Affy	Exploratory oligonucleotide array analysis.
8	affycomp	Compare expression measures for Affymetrix Oligonucleotide Arrays.
9	affyContam	Affymetrix cel file data
10	Affycoretools	Analyses with Affymetrix GeneChips
11	AffyExpress	Quality assessment and to identify differentially expressed genes in the Affymetrix affymetrix chip
12	Affyio	Parsing Affymetrix data files
13	affyImGUI	A Graphical User Interface for analysis of Affymetrix microarray gene expression data using the affy and limma Microarray packages
14	affyPara	Oligonucleotide array analysis
15	affyPLM	Quality assessment tools for affymetrix chip
16	affyQCReport	Quality of a set of affymetrix arrays
17	AffyRNADegradation	Assessment and correction of RNA degradation effects in Affymetrix 3' expression arrays
18	AffyTiling	Extraction and annotation of individual probes from Affymetrix tiling arrays.
19	annmap	Deep sequencing data analysis
20	arrayQualityMetrics	Microarray quality metrics reports
21	ArrayTools	Quality assessment and to detect differentially expressed genes for the Affymetrix GeneChips
22	ExpressionView	Isualizes possibly overlapping biclusters in a gene expression matrix.
23	limma	Linear Models for Microarray Data
24	MiChip	Microarray platform using locked oligonucleotides for the analysis of the expression of microRNAs in a variety of species
25	Simpleaffy	High level analysis of Affymetrix data
26	Starr	Affymetrix tiling arrays (ChIP-chip data)
27	yaqcaffy	Quality control of Affymetrix GeneChip expression data
28	annmap	Genome annotation and visualisation package pertaining to Affymetrix arrays

Table 1: Quality control packages of bio-conductor.

S. No	Name of Package	Description
1	DEGseq	To identify differential expressed genes from rna-seq data.
2	DEGreport	Report of deg analysis.
3	DEGraph	For two-sample test on graph.
4	DEDS	Differential expression via distance summary for microarray data.
5	DEseq2	Differential expression analysis based on negative binomial distribution.
6	DEXseq	For the inference of differential exon usage in rna-seq.
7	Dexus	For identifying the differential expression in rna-seq studies with unknown conditions or without replicates.
8	Derfinder	Used for annotation-agnostic differential expression analysis of rna-seq data at base-pair resolution.
9	Derfinder plot	To find the plot for derfind.
10	Diffbind	Used for differential binding analysis of chip-seq peak data.
11	Diffgeneanalysis	Perform differential gene expression analysis.

Table 2: Deg analysis packages of bio-conductor.

www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53890. GSE53890 is microarray experiment on REST and Stress Resistance in Aging and Alzheimer's disease (Figure 1) [22].

Result and Conclusion

From GSE53890 data out of 6 cel files are selected 3 females data and 3 males data chip used in this experiment is HG-U133_Plus_2. Gene expression file is created using affy package of bioconductor and visualization packages are used for quality control analysis and summarizing the output generated by these packages [23].

Gene expression file

Gene expression file created using affy package.

> library (affy)

> array = ReadAffy (widget=TRUE)

> eset = rma (array)

> write.exprs (eset, file = "array.txt")

Figure 2 shows the gene expression file and expression values. Using above commands in r this file can be created and used for analysis either in R or in any other softwares like MeV that not accept .cel and .cdf file as input. Either object can be created for expression file using >exprs=write.exprs (eset, file="array.txt")

Now this object exprs can be used directly in r to visualize expression file and analyze the intensity.

Visualization plots

There are number of plots built in quality control packages for visualization of chips and analysis these plots are also used for comparing chips before and after normalization. Here we describe some plots and their analysis.

Boxplot

Boxplot also called as box-and-whisker plot is statistical plot that graphically plots numerical values for comparison of chips. Five values are plotted in box plot that are lower min value, lower quartile, mean, upper quartile, max value of chip files in parallel for comparison of chip intensities. Box plot display variation in sample without assuming any statistical distribution. Spacing between box lines indicates the spread or degree of dispersion in intensities values of one sample. Boxplot can also be created for individual variables in R using Boxplot function and using fivenum command we can access the five values of expression (Figure 3).

> library (affyQCReport)

> affyQAReport (array)

Folder affyQA is created in same directory where r working directory is set. This folder contains individual graph files.

Intensity plot

Intensity plot is similar to Boxplot but it gives more detailed view. Intensity plot x-axis represents probe density and y-axis probe intensity. Figure 4 represents the intensity plot between 6 cel files any array whose intensity graph is very different from other array is considered as problematic.

RNA degradation plot

RNA degradation plot is used to assess the quality of RNA molecule used as probe in chips. Since RNA probes are designed from 3' end of mRNA molecule because RNA degradation starts from 5' end of mRNA molecule so intensity should be less in 5' end as compared to 3'end of probe. This plot is used for quality analysis of probes spotted on affymetrix chip (Figure 5).

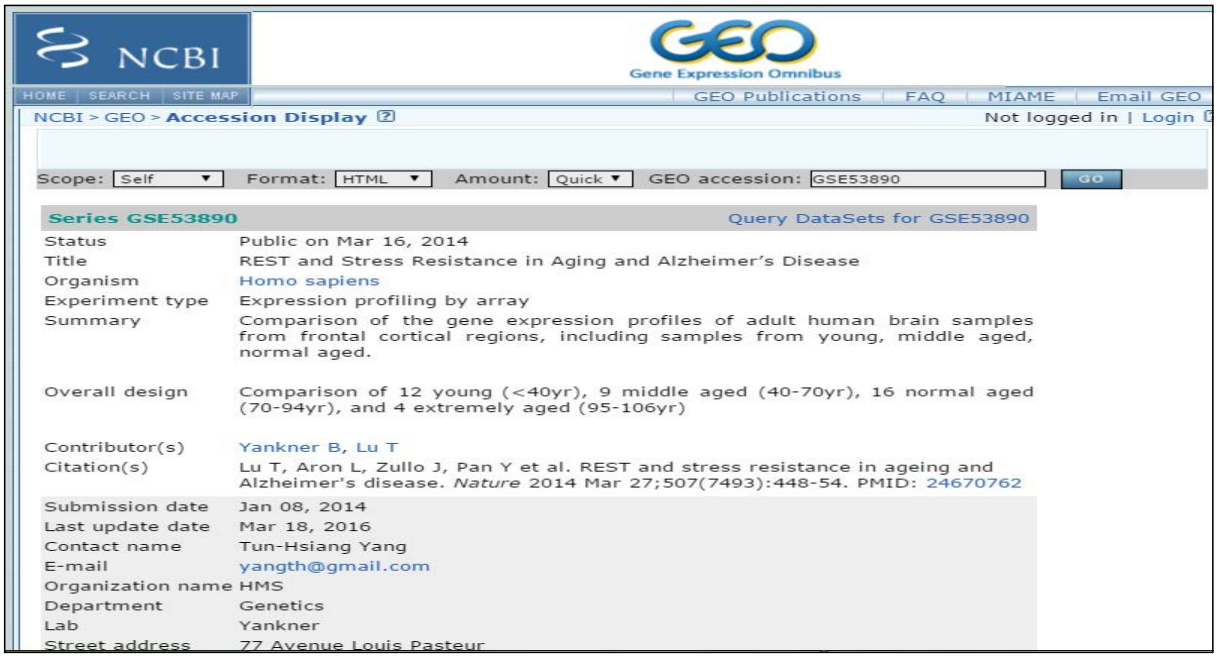

Figure 1: Gene expression Omnibus database.

	A	B	C	D	E	F	G	H
1		24 F.CEL	26 M.CEL	29 F.CEL	34 F.CEL	37 M.CEL	40 M.CEL	
2	1007_s_at	8.03429	8.772677	8.506786	8.806013	8.874959	8.471199	
3	1053_at	5.129433	5.26885	5.034688	5.031765	4.861491	4.917265	
4	117_at	4.944486	4.889163	4.710165	5.611009	4.827517	4.699408	
5	121_at	7.501474	7.355494	7.565741	7.450452	7.462104	7.602513	
6	1255_g_at	3.511665	3.580815	4.260613	3.928873	3.495896	3.578732	
7	1294_at	5.113248	5.258803	5.12377	4.937654	5.132642	5.137079	
8	1316_at	5.327564	5.341795	5.421229	5.412445	5.368434	5.43424	
9	1320_at	4.525333	4.445202	4.409657	4.461411	4.381585	4.396613	
10	1405_i_at	3.092261	2.880746	3.010714	3.030582	3.055263	3.075304	
11	1431_at	3.835525	3.752205	3.704723	4.03727	3.953105	3.936302	
12	1438_at	6.171631	5.769026	6.003033	5.839133	6.066168	5.910563	
13	1487_at	5.685933	5.872879	6.08433	6.100783	5.953952	6.146575	
14	1494_f_at	5.442937	5.490435	5.759878	5.539503	5.571146	5.545931	
15	1552256_a	6.993329	6.9624	6.836159	6.976911	7.118099	6.982972	
16	1552257_a	7.257667	7.068899	7.831783	7.312058	7.42301	7.59721	
17	1552258_a	3.781122	3.811404	3.816726	3.890403	3.699379	3.845689	
18	1552261_a	5.407248	4.970404	5.100554	5.080625	5.09663	5.054312	
19	1552263_a	2.968945	3.110073	3.036589	2.794043	3.069281	2.994463	
20	1552264_a	7.935421	7.499184	7.937578	7.256843	7.348546	6.982388	
21	1552266_a	3.375726	3.304053	3.527372	3.343631	3.184546	3.288351	
22	1552269_a	3.346049	3.267432	3.61881	3.190716	3.254016	3.181616	
23	1552271_a	4.683213	4.573254	4.539561	4.744894	4.597907	4.380566	
24	1552272_a	5.15859	4.690044	5.2258	4.978506	4.851615	5.034684	
25	1552274_a	5.627429	5.356057	6.119516	6.259067	6.417436	5.746686	

AGEING_exprs

Figure 2: The gene expression file and expression values.

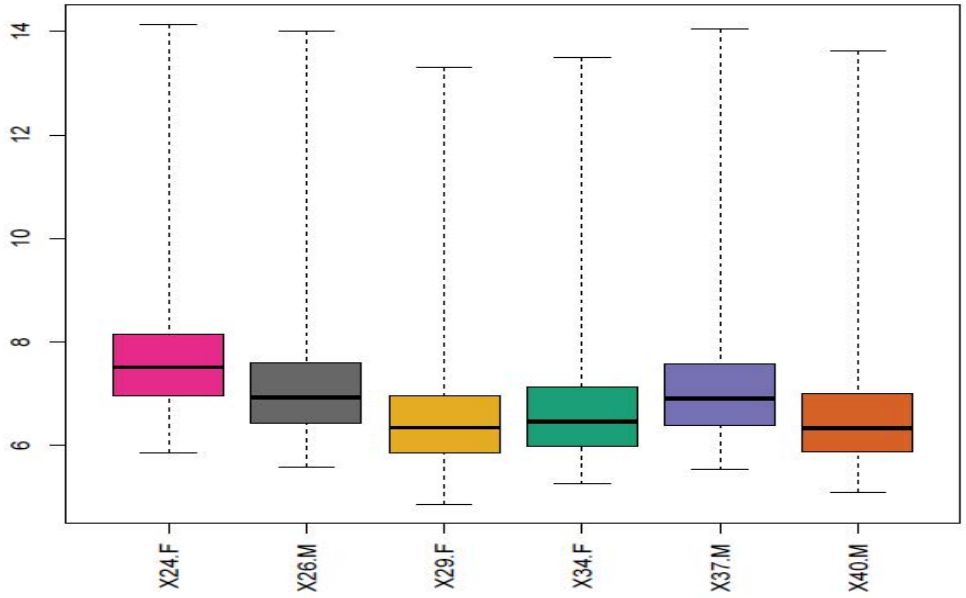

Figure 3: The Boxplot of gene expression file.

MA plot

MA plot where M represents minus sign and A represents mean addition sign for two channel microarray experiment.

Log Ratio: M ("Minus") = $\log_2(R/G) = \log_2 R - \log_2 G$

Average Log Intensity: A ("Add") = $\log_2 (RG)^{1/2}$ or $(1/2)(\log_2 R + \log_2 G)$

MA plot is used to determine is there any biasness in intensities of red and green dye and other systemic errors or instrumental errors in experiment. To visualize the need of normalization before any analysis.

MA plot is plot of red and green intensities biasness and determine the rate of error in microarray experiment. Figure 6 represents the ma plot between. cel files and variances in intensities of red and green dyes.

NUSE plot

Normalized Unscaled Standard Errors (NUSE) is used for standard error for each probe on all chips. All probes are normalized to scale one across all arrays. This plot shows the variations between genes and any array with higher standard error is of poor quality and hence can be rejected for further analysis for biological interpretations Figure 7 shows the NUSE plot.

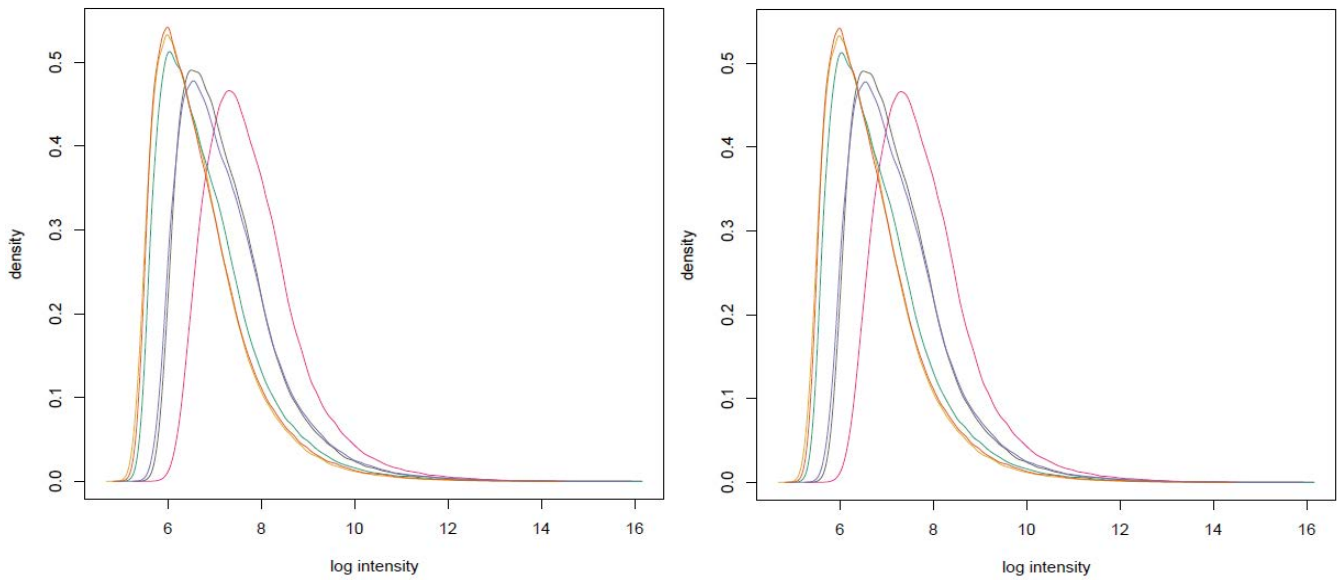

Figure 4: The intensity plot between 6 cel files any array.

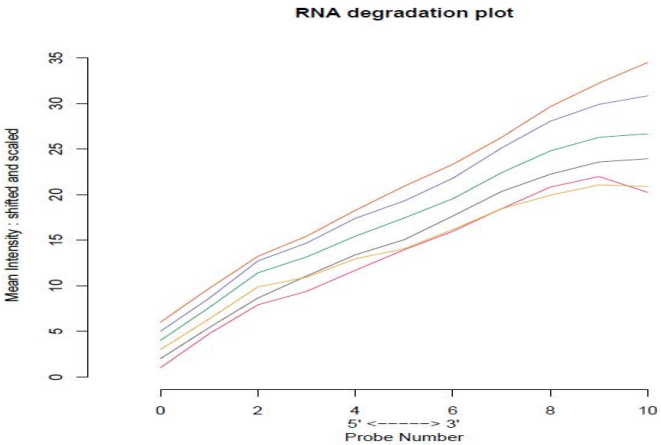

Figure 5: RNA degradation plot.

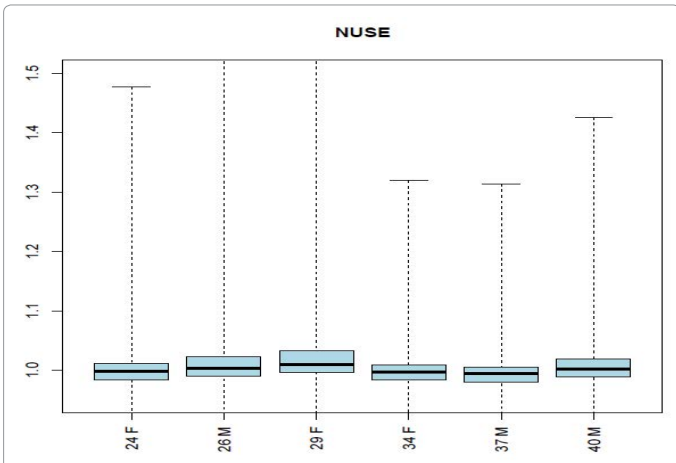

Figure 7: Nuse plot.

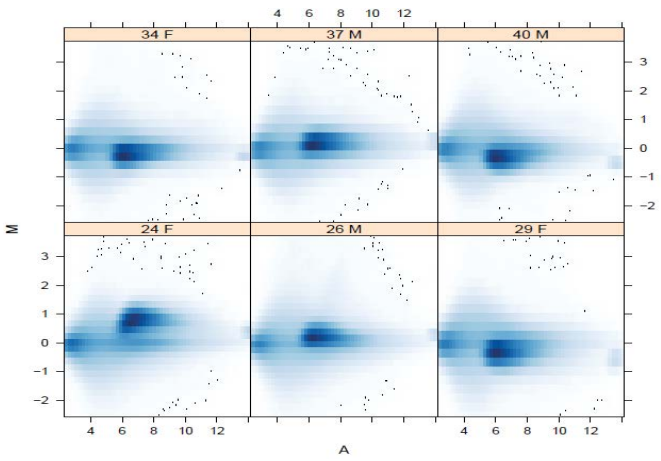

Figure 6: Ma plot between cel files and variances.

RLE plot

Relative Log Expression (RLE) plot is plot of RLE values that are calculated by assuming that all probeset across all array against median expression value for that probeset is constant and are not changing. Assuming that genes are constant across all arrays RLE values should be near 0.RLE plot is graphically plotted in form of Boxplot that provides quality of chips.

Figure 8 shows the RLE plot of chips in form of box plot and RLE values are set to 0. This plot shows the variation in chips and used for quality assessment of probes.

QC stats

QC plot is recommended by affymetrix. Any array that is shown in red is indication of error and in blue indicates the array within limits of scale factors e within 3-fold. Figure 9 represents the QC plot of arrays.

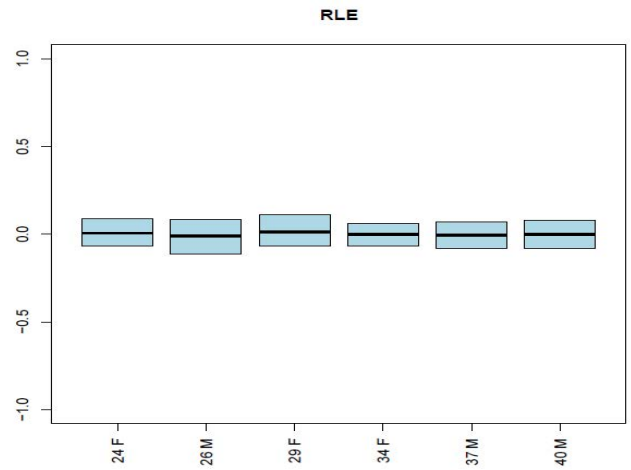

Figure 8: The variation in chips and used for quality assessment of probes.

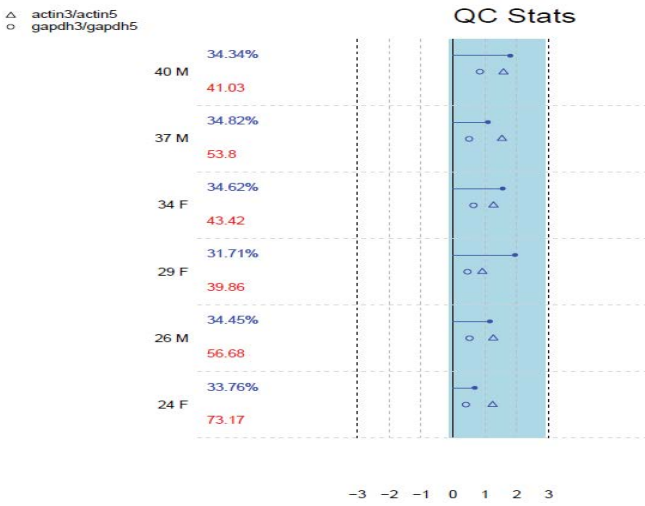

Figure 9: QC plot.

Conclusion and Discussion

This paper reviews the different quality control plots used for visualizing high throughput microarray data. And summarizes various packages of bioconductor that are used for microarray quality control and differentially expressed genes analysis. Any researcher working on microarray analysis can get help from this paper for microarray quality control analysis.

Acknowledgements

This is not just to follow the custom of writing acknowledgement but to express and record my heart felt feeling of thankfulness to all those who directly or indirectly helped me in this work. Further we wish to acknowledge bioinformatics tools for conducting this study.

References

1. Lu T, Aron L, Zullo J, Pan Y (2014) REST and stress resistance in ageing and Alzheimer's disease. Nature 507: 448-454.
2. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) Affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307-315.
3. Bolstad BM (2016) Affyio tools for parsing affymetrix data files. R package version 1: 40.
4. http://www.bioconductor.org,http://bioinformatics.picr.man.ac.uk/simpleaffy/
5. Parman C, Halling C, Gentleman R (2008) QC Report Generation for affy Batch objects. R package version 1.48.0.
6. Wu X, Li A (2009) ArrayTools: geneChip is Package. R package version 1.30.0.
7. Kauffmann A, Gentleman R, Huber W (2009) arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. Bioinformatics pp: 415-416.
8. Kauffmann A, Rayner TF, Parkinson H, Kapushesky M, Lukk M, et al. (2009) Importing ArrayExpress datasets into R/Bioconductor. Bioinformatics 25: 2092-2094.
9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43: 47.
10. Talloen W, Verbeke T (2011) a4: Automated Affymetrix Array Analysis Umbrella Package. R package version 1.18.0.
11. Bengtsson H, Bullard J, Hansen K (2015) affxparser: Affymetrix File Parsing SDK. R package version 1.42.0.
12. Talloen W, Verbeke T, Casneuf T, Bondt AD, Osselaer S, et al. (2013) a4Base: Automated Affymetrix Array Analysis Base Package.
13. Talloen W, Verbeke T (2011) a4Core: Automated Affymetrix Array Analysis Core Package.
14. Alloen W, Verbeke T (2011) a4Preproc: Automated Affymetrix Array Analysis Preprocessing Package.
15. Verbeke T (2010) a4Reporting: Automated Affymetrix Array Analysis Reporting Package.
16. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP (2004) affycomp: Graphics Toolbox for Assessment of Affymetrix Expression Measures.
17. Smith CA (2010) annaffy: Annotation tools for Affymetrix.
18. Biological metadata (2013) R package version 1.42.0.
19. Ates T (2011) Annmap: Genome annotation and visualisation package pertaining to Affymetrix arrays and NGS analysis.
20. Zacher B, Soeding J, Kuan PF, Siebert M, Tresch A (2010). Starr: Simple tiling array analysis of Affymetrix ChIP-chip data. BMC Bioinformatics 11: 194
21. http://www.ncbi.nlm.nih.gov/geo
22. https://www.bioconductor.org/
23. https://cran.r-project.org/