PAKISTAN BIOMEDICAL JOURNAL

https://www.pakistanbmj.com/journal/index.php/pbmj/index

Volume 6, Issue 4 (April 2023)

Original Article

Comparative Analysis of R and Mathematica Package for Differential Gene Expression Analysis Using Microarray Dataset on Pancreatic Cancer

Kinza Qazi[1*] and Tehreem Anwar[2]

[1]University of Veterinary and Animal sciences, Lahore, Pakistan
[2]Virtual University of Pakistan, Lahore, Pakistan

ARTICLE INFO

Key Words:

R Language, Mathematica, Differential Gene Expression, Microarray

How to Cite:

Qazi, K. ., & Anwar, T. . (2023). Comparative Analysis of R and Mathematica Package for Differential Gene Expression Analysis Using Microarray Dataset on Pancreatic Cancer: Comparative Analysis of R and Mathematica Package. Pakistan BioMedical Journal, 6(04). https://doi.org/10.54393/pbmj.v6i04.863

***Corresponding Author:**

Kinza Qazi1*
University of Veterinary and Animal sciences, Lahore, Pakistan
kinzaqazi04@gmail.com

Received Date: 27th March, 2023
Acceptance Date: 22nd April, 2023
Published Date: 30th April, 2023

ABSTRACT

Microarrays produces enormous amounts of information requiring a series of repeated analyses to condense data. To analyze this data several computational software is used. **Objective:** To compare the analysis of R and Mathematica package for differential gene expression analysis using microarray dataset. **Methods:** Microarray Data were collected from an online database GEO (gene expression omnibus). Mathematica and R software was used for comparative analysis. In R software, Robust Multi-Array Average (RMA), was used for data normalization. While Limma package was used for DGE analysis. In Mathematica software, AffyDGED was used for normalization and DGE analysis of dataset. **Results:** 3,426 non-differentially expressed genes and 14936 genes with differential expression were separated from R. The thresholds for identifying "up" and "down" gene expression were estimated to be 0.98 and -0.19, respectively, using the RMA method to analyze this dataset. AffyDGED from Mathematica detected 1,832 genes as differentially expressed; of them, 1,591 genes overlap with the real and 1,944 differently expressed genes, giving the true positive rate of (1591/1944) =0.818. This indicates that 18% of the genuine list of differentially expressed genes could not be reliably identified by AffyDGED. **Conclusions:** R programming is one of the most popular and recommendable tools for microarrays to perform different analysis, and along with Bioconductor it makes one of the best analysis algorithms for DGE analysis. On the other hand, AffyDGED brings a contemporary algorithm useful in the real world to the Mathematica user.

INTRODUCTION

Pancreatic cancer is the fourth top reason of demise caused by cancer [1]. Lack of early detection technology for pancreatic cancer invariably leads to a typical clinical appearance of incurable disease at initial diagnosis [2]. Pancreatic cancer originates when glandular organ behind stomach starts an abnormal growth in pancreatic cells and goes out of control to form a mass structure [3]. Microarray is a technology that concurrently estimates the quantitative measurements for the expression of thousands of genes [4]. A microarray is a complex commonly known as a lab-on-a-chip [5]. A (2 Dimensional) array on a solid substrate like glass slide or silicon thin cell which appraises excessive amounts of biological material using high-throughput screening reduced for parallel dispensation and detection methods [6]. To evaluate gene expression between clusters of cells of different organs or individual's DNA microarrays is used [7]. Gene expression analysis is a method in which information from gene is used to synthesize a valuable gene product. In most cases, these products are functional proteins but, in some cases non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes are synthesized in form of functional RNA [8]. The process consists of few steps which includes, transcription, RNA splicing, translation, post translational modification and gene regulation. Gene regulation is operated by a cell regulator for structure and function which originates from cellular differentiation, morphogenesis, versatility, and adaptability of any

DOI: https://doi.org/10.54393/pbmj.v6i04.863

organism [9]. It also assists as a substrate for evolutionary change [10]. Bioinformatics become a progressively significant tool for molecular biologists, specifically for the analysis of microarray data. Bioinformatics collaborates with different computational tools which incorporates with analysis of biological and medical data [11]. Microarrays produces enormous amounts of information requiring a series of repeated analyses to condense data [12]. Through output of microarrays direct interpretation is not possible to show differences in conditions of samples, or time points. To create microarray experiments interpretable, it requires a sequence of algorithms and methods to applied [13]. After normalization of generated data, which is required to make a contrast possible, significance analysis, clustering of samples and biological composites of interest and visualization are usually achieved. Microarray experiments deals with several bioinformatics challenges [14]. R is free and open-source Platform not only for computational analysis, but also very useful in the field of bioinformatics and their related analysis such as, Gene expression analysis, Gene knockout findings, Microarray data analysis [15]. Mathematica is a computer algebra system or program, used mostly in calculating fields of applied mathematics professionals [16]. Major Objective of this study was to analyze microarray data of pancreatic cancers created by using HG-U133_Plus_2 Affymetrix Human Genome U133 Plus 2.0 Array platform. By using this data with different programming languages and their acquired packages a comparison of expression analysis was done and different genes was discovered on the basis of their regulatory effect. This research will further enlighten a path for many different statistical studies in the field of biological and computational biological data sciences. It will also enable new ways to look forward towards the fields of personalized and predictive medicine which will encourage scientists to develop new therapeutic advancements to control chronic genetic diseases.

METHODS

This study analyzed the microarray data of healthy vs pancreatic cancer patient trough R language and Mathematica software. Total 24 samples of 24 different patients were used in this study from which 12 samples were taken from pancreatic cancer patients and 12 from normal healthy patient. Microarray Data were collected from an online database GEO (gene expression omnibus). The dataset which we took from GEO had Accession number GSE14245. This poised dataset was built on transcriptomic approach that profiled the saliva samples from 12 pancreatic cancer patients and 12 healthy control subjects using the Affymetrix Human Genome U133 Plus 2.0 Array platform [17]. R Language and Mathematica was used for the analysis. First step was data normalization followed by differential gene expression (DGE) analysis and extraction of up and down regulated genes. Data normalization is done to remove any zero or negative counts to make data less contaminated and easy to use for further analysis. In this research, our major focus was on the comparison of algorithms used for normalization and DGE analysis in both platform (R and Mathematica). In R software, Robust Multi-array Average (RMA), was used for data normalization. While Limma package was used for DGE analysis. In Mathematica software, AffyDGED was used for normalization and DGE analysis of dataset. Detailed methodology is shown in Figure 1.

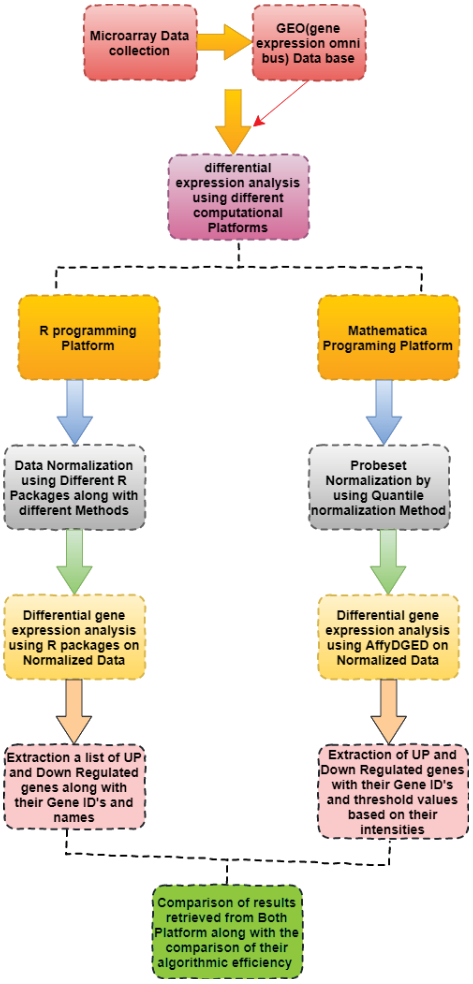

Microarray Data collection
GEO(gene expression omni bus) Data base
differential expression analysis using different computational Platforms
R programming Platform
Mathematica Programing Platform
Data Normalization using Different R Packages along with different Methods
Probeset Normalization by using Quantile normalization Method
Differential gene expression analysis using R packages on Normalized Data
Differential gene expression analysis using AffyDGED on Normalized Data
Extraction a list of UP and Down Regulated genes along with their Gene ID's and names
Extraction of UP and Down Regulated genes with their Gene ID's and threshold values based on their intensities
Comparison of results retrieved from Both Platform along with the comparison of their algorithmic efficiency

Figure 1: Overall methodology of the study

RESULTS

In Figure 1, we can see the difference between boxplots of log-intensity distribution which are plotted to check the difference between distribution. After RMA normalization we can easily see a comparable difference. Specially if we see sample no. 21 in both figures, there is some visible difference that means several zero counts has been removed from this dataset after normalization.

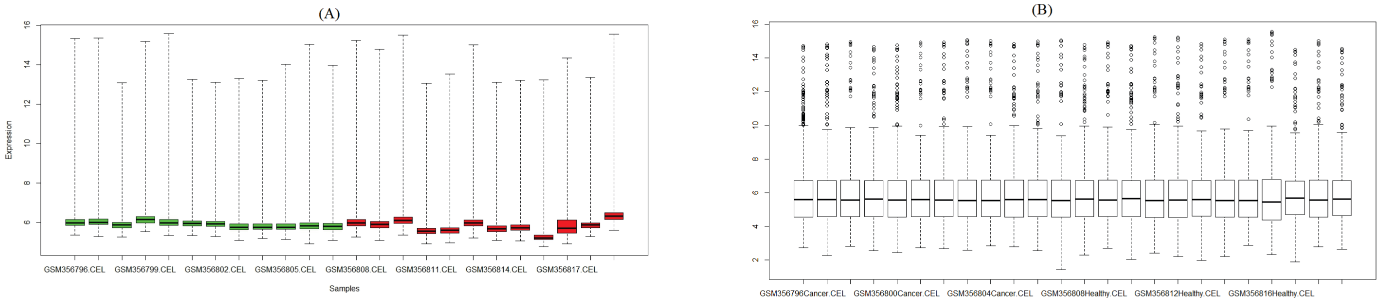


Figure 2: (A) Box plot showing probe intensity of different genes present in microarray dataset (B) Microarray data variation after normalization using RMA method

After normalization, DGE analysis was performed by R software using Limma package. The data of up and down regulated gene extracted from dataset is mentioned in Table 1. Almost 6.2% of genes in this data did not show any regulatory effects in any case and almost 67% genes were down regulated which means that the effect of genes having low expression values doesn't show much possibility to trigger disease (Table 1).

Table 1: Results of UP and DOWN regulated genes

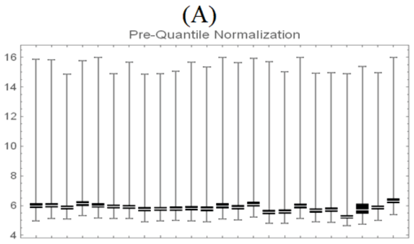
Trend	Genes
UP-Regulated genes 0 < -1	14936
Not-affected genes 0	3426
Down-Regulated genes > 1	36313

Table 2 shows a list of 8 up and down regulated genes based on the highest logFC and p-values. The estimated logFC for multiple treatment conditions compared back to the same control group will be positively correlated even in the absence of any biological effect. Maximum values in up regulated genes estimated by R were 4.8 and minimum value in up regulated expression was 0.98. In down regulation, maximum value estimated was -11.23 and minimum value was -0.19.

Table 2: Top 8 values from up and down regulated differential gene expressions

ID	logFC	Avg. Expression	T	p-value	adj. p-value	B	Fold Change Cancer/Normal	Gene symbol
241174_at	11.44353	11.52234	48.80763	5.02E-27	2.75E-22	40.76645	2785.1382	AP4E1
1553088_a_at	10.40041	10.43073	44.01448	7.09E-26	7.73E-22	39.72194	1351.55932	BCL2L11
1552899_at	9.726866	9.696013	43.70573	8.48E-26	7.73E-22	39.64606	847.3806098	LINC01312
1557866_at	10.1436	10.33103	43.18777	1.15E-25	8.99E-22	39.51616	1131.169839	CFAP157
243269_s_at	9.989396	10.03627	42.43105	1.81E-25	1.15E-21	39.32031	1016.501089	FAM205BP///FAM205A
230092_at	9.465539	9.446353	42.25046	2.02E-25	1.15E-21	39.27247	706.9864018	UBXN10
1564662_at	9.252544	9.308898	41.59133	3.01E-25	1.37E-21	39.09416	609.9485222	ZNF852
208191_x_at	10.6127	10.64971	41.23602	3.75E-25	1.58E-21	38.99557	1565.812265	PSG4

Figure 3 shows pre- and post-quantile normalization done by mathematics software. In pre-quantile plot we use raw data as an input to make these boxplots. Here we clearly see variation among samples because there is no elimination of perfect match probe intensities are applied on it. Form sample no.14 up till sample no. 24 all these samples are from normal patients but shows a large amount of variation amongst all these.



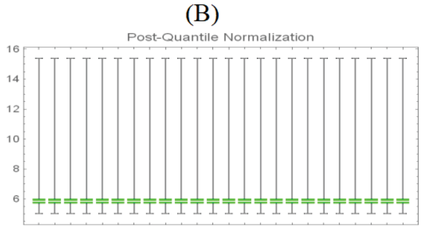


Figure 3: (A) A box-and-whisker comparison before quantile normalization performed by Mathematica (B) After quantile normalization results

In Figure 4 below, X-axis show threshold data of DE that is total number of genes present in an entire dataset which was almost 54130 genes as per Mathematica AffyDGED algorithm analyzed. While on Y-axis values of up and down regulated threshold is presented. In this case, our minimum

DOI: https://doi.org/10.54393/pbmj.v6i04.863

cutoff threshold value of up regulated genes is 0.18 and maximum value is 0.49 whereas minimum value of down regulated genes is -0.155 and maximum value is -0.78.

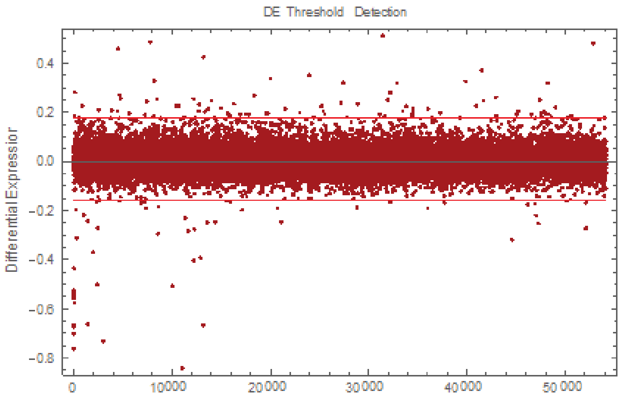


Figure 4: Image showing UP and Down regulated DE genes threshold detection

Table 3 shows top 15 up regulated genes extracted using Mathematica, the best part in AffyDGED algorithm is that it also gives a very detail view of comparison between control and experimental group and also give the values for how genes are differentially expressed in experimental group which is known as cutoff value.

Table 3: Up regulated genes extracted by Mathematica

Affymetrix probe set name	fluorescence intensity of GE in the experimental group	fluorescence intensity of GE in the control group	p-value	Cutoff values of differential gene expression	Gene symbol	GenBank. Accession
241174_at	11.44353	11.52234	5.02E-27	2.75E-22	AP4E1	AV647279
1553088_a_at	10.40041	10.43073	7.09E-26	7.73E-22	BCL2	NM_138626
1552899_at	9.726866	9.696013	8.48E-26	7.73E-22	L11LIN	-
1557866_at	10.1436	10.33103	1.15E-25	8.99E-22	C01312C	AK094948
243269_s_at	9.989396	10.03627	1.81E-25	1.15E-21	FAP157FAM205BP///FAM205A	AL040346
230092_at	9.465539	9.446353	2.02E-25	1.15E-21	UBXN10	AA135547
1564662_at	9.252544	9.308898	3.01E-25	1.37E-21	ZNF852	BC014381
208191_x_at	10.6127	10.64971	3.75E-25	1.58E-21	PSG4	NM_002780
215856_at	9.105225	9.173238	4.30E-25	1.68E-21	SIGLEC15	AK025833
242316_at	9.097469	9.181627	6.84E-25	1.82E-21	TMOD3	AI810103
208257_x_at	10.64889	10.67999	7.48E-25	1.82E-21	PSG1SHI	NM_006905
1556619_at	9.284751	9.340856	7.75E-25	1.82E-21	SA9	CA413715
226611_s_at	9.76453	9.899672	7.94E-25	1.82E-21	CENPV	AA722878
208469_s_at	9.338312	9.495589	8.01E-25	1.82E-21	PPT2-EGFL8///EGFL8///PPT2	NM_030652

DISCUSSION

This thesis was basically focused on using bioinformatics techniques i.e., statistical computing and algorithms to analyze datasets and perform differential expression analysis on it. The Pancreatic cancer Microarray dataset (GSE14245) was used, which was extracted by using the transcriptomic approach profiled the saliva supernatant samples from 12 pancreatic cancer patients and 12 healthy control subjects using the Affymetrix Human Genome U133 Plus 2.0 Array platform. There are very few research articles are available which shows the difference between both of these tools. Not only in tools and platform our interest is also towards the efficiency of algorithms used by these platforms to preform differential expression analysis. RMA is one of the most commonly used algorithms which give normalized data after eliminating the mismatch probe values so it gives a good quality of normalized values moreover, it also has a quantile normalization method which compare background correction within each probeset ratio. Irizarry *et al.*, also performed similar analysis on dataset generated from Affymetrix GeneChip system. The dataset was of high-density oligonucleotide array data. They explained why there is a need to examine and normalize microarrays datasets using probe level densities. They also used RMA algorithm for normalization, and they concluded that there was no shortcoming for using RMA for normalization of microarray data [18]. Mathematica used AffyDGED algorithm for DE analysis which is somewhat similar to RMA but have a lot of differences as well, so we can say that AffyDGED is a mixture of both mas5 and RMA. When results were

DOI: https://doi.org/10.54393/pbmj.v6i04.863

retrieved from Mathematica, our next step was comparison. For this we take average scores of up and down regulated gene expression and also check how many genes are up and down regulated and also compare the amount of not effected or overlapped genes. 3,426 non-differentially expressed genes and 14936 genes with differential expression were separated from R. The thresholds for identifying "up" and "down" gene expression were estimated to be 0.98 and -0.19, respectively, using the RMA method to analyze this dataset. A plot of the processed data always reveals a tight clustering of data about the line y=0. This observation was used to develop code that scans vertically up and down in small increments and establishes a breakpoint in each direction any time the density of data at a vertical position is 50% less than it was at the previous increment. These breakpoints become the thresholds for determining differentially expressed up and down genes. Gregory Alvord *et al.*, performed DEG analysis of microarray data from Soybean genome. They also used Rand Bioconductor for DEG analysis and RMA algorithm for normalization. These programs successfully identified differential gene expression results from soybean genome data [19]. AffyDGED from Mathematica detected 1,832 genes as differentially expressed; of them, 1,591 genes overlap with the real and 1,944 differently expressed genes, giving the true positive rate of (1591/1944) =0.818. This indicates that 18% of the genuine list of differentially expressed genes could not be reliably identified by AffyDGED. Allen also studied differential gene expression using Affymetrix microarrays. He also used AffyDGED algorithm of Mathematica for analysis. AffyDGED algorithm performed very well and took very less time for analysis [20].

CONCLUSIONS

Microarray technology continues to be heavily used by the biomedical and basic science research communities throughout the world. R programming is one of the most popular and recommendable tools for microarrays to preform different analysis, and along with Bioconductor it makes one of the best analysis algorithms for DGE analysis. On the other hand, AffyDGED brings a contemporary algorithm useful in the real world to the Mathematica user, but this is not much familiar to every researcher so, it is much needed to explore this software by those who have interest in exploring fundamental biology questions with their favorite computational tool chest.

Conflicts of Interest

The authors declare no conflict of interest.

Source of Funding

The authors received no financial support for the research, authorship and/or publication of this article.

REFERENCES

[1] Michaud D. Epidemiology of pancreatic cancer. Minerva Chirurgica. 2004 Apr; 59(2): 99-111.

[2] Kaur S, Baine MJ, Jain M, Sasson AR, Batra SK. Early diagnosis of pancreatic cancer: challenges and new developments. Biomarkers in Medicine. 2012 Oct; 6(5): 597-612. doi: 10.2217/bmm.12.69.

[3] Yang S, Wang X, Contino G, Liesa M, Sahin E, Ying H, *et al*. Pancreatic cancers require autophagy for tumor growth. Genes & Development. 2011 Apr; 25(7): 717-29. doi: 10.1101/gad.2016111.

[4] Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S, Bennett H, *et al*. A novel, high-performance random array platform for quantitative gene expression profiling. Genome Research. 2004 Nov; 14(11): 2347-56. doi: 10.1101/gr.2739104.

[5] Mark D, Haeberle S, Roth G, Von Stetten F, Zengerle R. Microfluidic lab-on-a-chip platforms: requirements, characteristics and applications. Chemical Society Reviews. 2010 Jan; 39: 1153-82. doi: 10.1039/b8205 57b.

[6] Heller MJ. DNA microarray technology: devices, systems, and applications. Annual Review of Biomedical Engineering. 2002 Aug; 4(1): 129-53. doi: 10.1146/annurev.bioeng.4.020702.153438.

[7] Higgins JP, Shinghal R, Gill H, Reese JH, Terris M, Cohen RJ, *et al*. Gene expression patterns in renal cell carcinoma assessed by complementary DNA microarray. The American Journal of Pathology. 2003 Mar; 162(3): 925-32. doi: 10.1016/S0002-9440(10)638 87-4.

[8] Lockhart DJ and Winzeler EA. Genomics, gene expression and DNA arrays. Nature. 2000 Jun; 405(6788): 827-36. doi: 10.1038/35015701.

[9] Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PloS One. 2017 Dec; 12(12): e0190152. doi: 10.1371/journal.pone.0190152.

[10] Márquez-Zacarías P, Pineau RM, Gomez M, Veliz-Cuba A, Murrugarra D, Ratcliff WC, *et al*. Evolution of cellular differentiation: from hypotheses to models. Trends in Ecology & Evolution. 2021 Jan; 36(1): 49-60. doi: 10.1016/j.tree.2020.07.013.

[11] Can T. Introduction to Bioinformatics. Methods in Molecular Biology. 2013 Nov; 1107: 51-71. doi: 10.1007/978-1-62703-748-8_4.

[12] Harrington CA, Rosenow C, Retief J. Monitoring gene expression using DNA microarrays. Current Opinion in Microbiology. 2000 Jun; 3(3): 285-91. doi: 10.1016/S1369-5274(00)00091-6.

[13] Kerr MK and Churchill GA. Statistical design and the analysis of gene expression microarray data.

Genetics Research. 2001 Feb; 77(2): 123-8. doi: 10.1017/S0016672301005055.

[14] Xu J, Shu Y, Xu T, Zhu W, Qiu T, Li J, *et al*. Microarray expression profiling and bioinformatics analysis of circular RNA expression in lung squamous cell carcinoma. American Journal of Translational Research. 2018 Mar; 10(3): 771-83.

[15] Chambers JM. Software for data analysis: programming with R. New York: Springer; 2008. doi: 10.1007/978-0-387-75936-4.

[16] Maeder R. Programming in mathematica. Addison-Wesley Longman Publishing Co., Inc.; 1991.

[17] Afshari CA, Nuwaysir EF, Barrett JC. Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. Cancer Research. 1999 Oct; 59(19): 4759-60.

[18] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr; 4(2): 249-64. doi: 10.1093/biostatistics/4.2.249.

[19] Gregory Alvord W, Roayaei JA, Quiñones OA, Schneider KT. A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. Briefings in Bioinformatics. 2007 Nov; 8(6): 415-31. doi: 10.1093/bib/bbm043.

[20] Allen T. Detecting differential gene expression using affymetrix microarrays. The Mathematica Journal. 2013; 15: 1-26. doi: 10.3888/tmj.15-11.