https://doi.org/10.48047/AFJBS.5.3.2023.119-130

African Journal of Biological Sciences

Research Paper Open Access

Lung cancer data analysis for finding gene expression

Arshed H.Yaseen[1,a], Afrah Toama Khalaf [2] and Mohammed Ahmed Mustafa[3,b]

[1-3]Lect. Department of Biotechnology, College of applied Science, University of Samarra. Samarra, Iraq.

[2] Assist. Prof. Department of Biology, College of Education, University of Samarra, Iraq.

3-Department of medical Laboratory technology, Imam Jaafar Al-Sadiq university

email: [a] arshed.h87@uosamarra.edu.iq

[b] mohammed.alsad3@gmail.com

Article Info

Volume 5, Issue 3, July 2023

Received:18 May 2023

Accepted :15 June 2023

Published: 27 July 2023

doi:10.48047/AFJBS.5.3.2023.119-130

Abstract

Lung cancer harmful tumors originating from aviation route epithelioma, is the regularly analyzed disease on the planet and the most successive reason for malignancy passing. Transcriptomics has recently become a useful tool in systems biology, thanks to the use of DNA microarrays. Despite the availability of numerous types of microarrays today, including DNA, antibody, and protein microarrays, the term "microarray" will be used in this paper to refer to DNA microarrays only unless otherwise stated. The data on lung cancer in humans was acquired from the CEO data set and analyzed with R software. The LIMMA package in R was used to find a set of significantly variable genes, which are listed in Table-. The list contains 1053 genes (550 of which are up-regulated and 503 of which are down-regulated), all of which have a p-value less than 0.05. A bunch of fundamentally fluctuating qualities was discovered using the LIMMA package in R and displayed in table-. There are more than 1000 genes in the rundown (600 as up-regulated and more than 500 as down-regulated), recognized by p value less than 0.05.

Keywords: Lung cancer, Package, R program, Annotation

Introduction

Cancer of the lungs, i.e., bronchogenic harmful tumors originating from aviation route epithelioma, is the regularly analyzed disease on the planet and the most successive reason for malignancy related deaths. Roughly 1.8 million new cases of cellular breakdown in the lungs are analyzed around the world. In 2012, around 1.6 million individuals died from lung cancer and the number is estimated to exceed 3 million by 2035 (Didkowska el al., 2016). Cellular breakdown in the lungs has a moderately helpless anticipation, and the 5-year endurance differs from 4% to 17%, contingent upon the phase of the infection at the hour of its conclusion (Hirsch, F.R, at el, 2017). The headway of non-obtrusive diagnostics improves the chance of identifying cellular breakdown in the lungs, but be that as it may, just 10–15% of new cases are analyzed at its clinical phase (Xi, at el, 2018). About 75% of patients are determined to have cellular breakdown in the lungs at its high-level stage, when treatment choices are restricted. All things considered, in patients with clinical stage IA illness in the TNM (tumor-lymph hubs metastasis) arrangement, the 5-year endurance comes to roughly 60%, which demonstrates that countless patients suffer imperceptible metastases at this phase of the illness (Lu et al. 2018). Tools for the investigation of biological systems have lately become more powerful in terms of both experimental efficiency and the volume of information produced, tools for investigating biological systems have lately gotten significantly more powerful. This data can be utilized to create biological system models that anticipate the outcome of specific inputs. Traditionally, data has been collected in a time-consuming manner by assessing the effects of changes on a single or a small number of targets. High-throughput approaches in fields like transcriptomics, metabolomics, and proteomics are currently transforming this process, revealing previously unknown genetic and regulatory linkages. Alex et al, 2011). Transcriptomics is concerned with the transcription of a cell's messenger RNA (mRNA), while proteomics is concerned with the cell's protein complement; metabolomics is concerned with systematic variations in the concentration of metabolites.

Transcriptomics has become a useful tool in systems biology thanks to the use of DNA microarrays (Andrew et al 2010). Despite the availability of numerous types of microarrays today, including DNA, antibody, and protein microarrays, the term "microarray" will be used in this paper to refer to DNA microarrays only unless otherwise stated.

DNA microarrays Technique supply of a wide view in all direction and quantitative overview of a samples to get result of an expression. The force is massive as biological processes overall output from the coordinated interfere of multiple-genes (Andrew et al., 2010).

The 'transcriptome,' or the whole complement of cellular mRNA transcripts can be measured using DNA microarrays, providing a cue of the overall gene activity at any given time. Rather than focusing on the interconnections, expression or regulation of a few genes,

microarrays allow researchers to simultaneously examine overall gene expression, allowing them to infer associations and interactions. These measurements are used as the basis for the detection of disease processes, basic cellular metabolism, and cell cycle regulation.

Methods and Materials

1- pipeline data By RStudio Software v. 1.2.5033

Raw data → Normalization → Gene expression (statistics multiple testing) → up and down regulated → Annotation → Ontology and pathways.

1.1- Importing of Affymetrix data

Affyymetrix data is generated via the processing of a .DAT image file to produce a .CEL file that contains a single number that defines its intensity for each probe on the array. Package (affy) offers the basic techniques for the assessment and analysis of affymetrix oligonucleotide arrays.

1.2- Normalizing Affymetrix data

Normalization is simply one step in Affymetrix data processing that must be completed before gene expression estimations can be used for further analysis. Background correction, normalizing of probes, and summarization (where individual probes are integrated into a probeset) are typical processes in preprocessing methods like RMA. We'll use RMA preprocessing in this study to clean up the data. RMA was chosen due to observations that it provides highly exact expression estimates (which is needed), albeit it may not produce as accurate findings as MAS5. RMA, in other words, appears to consistently underestimate gene expression.

1.3- Plotting, clustering and quality control

The raw data (Affymetrix arrays) are subjected to quality control for . The Affymetrix CEL files were read as raw data in an object. That raw data can be used to do quality checks. The basic parts of the Affymetrix's basic quality control include checking for RNA degradation and assessing the expression of scaling factors, control genes, % of present genes, and the average background. The functions used to conduct these analyses are split across two packages- (affy) and (affy) (simpleaffy). Additionally, non-metric multidimensional scaling, Scatter plot, hierarchical clustering, boxplot, MA plot, and boxplot can be used to supplement these fundamental techniques.

1.4- Filtering tools and differential gene expression

Unspecific filtering is commonly used to exclude genes that are uninteresting from a dataset. Genes that do not change during the experiment or those that are expressed at such a low level that their measurements are inaccurate, are usually left out of future

analysis. We're utilizing two common unspecific filtering methods in this work; one based on expression and the other on standard deviation. Genefilter is a library with ready-to-use filtering capabilities. We generate a standard deviation for each gene before applying the standard deviation filter. Filtering is done at multiple stages; first, the function is used to generate a filtering function F. Then, using an auxiliary function called genefilter, this function is applied to all rows of the matrix (). We'll assume that we're looking for 1.5-fold overexpression (A=0.75) and that the gene must be overexpressed in not less than 50% of the arrays (p=0.05).

1.5- Statistical analysis

A common assumption of the Limma's method is that the data is normally distributed (else, the significance tests will yield incorrect findings), yet data is not always normally distributed in the actual world. Only about 20% of the expression values in a typical Affymetrix experiment are normally distributed. Others are non-normally distributed, and non-parametric approaches should be used to analyze them. Linear models are extremely adaptable tools for assessing even the most complex experimental sets.

1.6- Gene set analysis

Typically, hypergeometric test-based statistics are used in these procedures. Gene set enrichment analyses are approaches in which genes are initially assigned to categories or pathways and their statistical relevance is examined utilizing both the category information and the expression data. To determine the statistically substantially regulated genes from the data, a statistical procedure called gene set enrichment analysis is used.

1.7- The list of genes annotation

A vector of gene names is used as an input in the annotation process. These are usually taken from a Limma result matrix. The method produces a text or HTML file with the annotations as its output. Package (annaffy) implements the actual annotation process. This package, despite its name, can also be used to annotate other chip types as long as they have a suitable annotation package.

Results and Discussion

1- read affymatrix data

From lung cancer CEL files, we found the number of samples (36) and total number of lung cancer genes (54675) while the package annotation is (hgu133plus2) Figure (1)

```
> raw.data
AffyBatch object
size of arrays=1164x1164 features (22 kb)
cdf=HG-U133_Plus_2 (54675 affyids)
number of samples=36
number of genes=54675
annotation=hgu133plus2
notes=
```

Fig1: - CEL file output

2- Normalization

It uses the RMA approach to generate normalized and background adjusted expression values. The resulting data is saved as (ExpressionSetclass). The more memory-efficient justRMA() function is used for large data sets. The data is saved after normalization in a format specific to Bioconductor tools, such as affy or beadarray (Figure 2).

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 54675 features, 36 samples
  element names: exprs
protocolData
  sampleNames: GSM4504101_SCLC_01_ca.CEL.gz GSM4504102_SCLC_01_n.CEL.gz
    ... GSM4504136_SCLC_20_n.CEL.gz (36 total)
  varLabels: ScanDate
  varMetadata: labelDescription
phenoData
  sampleNames: GSM4504101_SCLC_01_ca.CEL.gz GSM4504102_SCLC_01_n.CEL.gz
    ... GSM4504136_SCLC_20_n.CEL.gz (36 total)
  varLabels: sample
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation: hgu133plus2
```

Fig (2): - Normalization expression value

3- Plotting, clustering and QC.

Different chips are separated by vertical grey lines in the qc plot. The red figures on the left indicate the number of probesets, with the present flag and the chip's average backdrop. The blue central region denotes the area where the scaling factors are less than three times those of all chips combined. The scaling factors for the chips are indicated by bars that end in a point. The open circles represent GADPH 3':5' ratios, while the triangles represent beta-actin 3':5' ratios. Not all of the scaling factors are within the permitted range in this case (Figure 3). While plotting, we used a Box plot to compare the samples before and after normalization using the RMA approach (Figure 4). To create a dendrogram for the hierarchical clustering, the first step is to use the command dist () to calculate all the pairwise distances between the samples. The unweighted pair group technique with arithmetic mean (UPGMA or average linkage) method is then used to put these distances between data into a dendrogram. UPGMA is the default method for tree classification in the command hclust() (Figure 5).

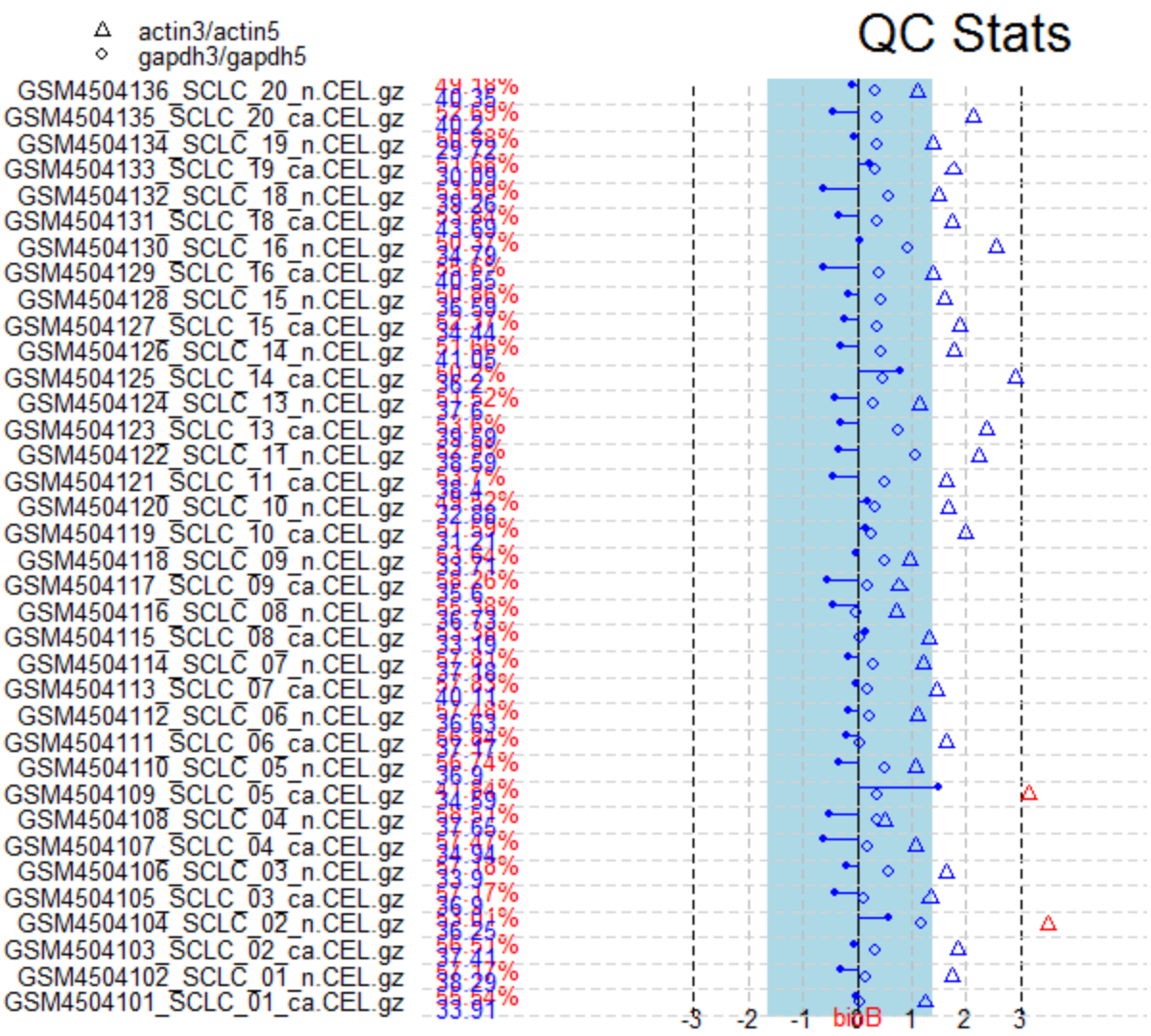
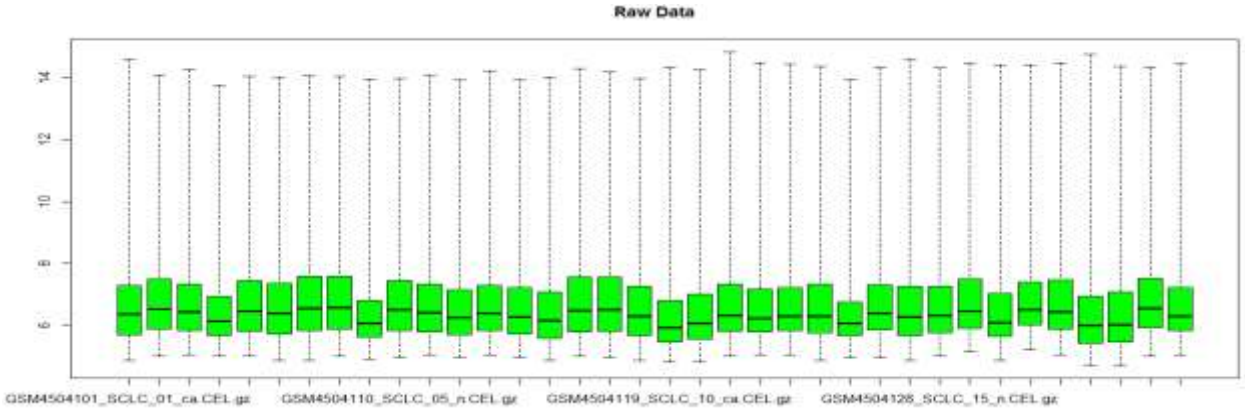


Fig (3): - qc state result



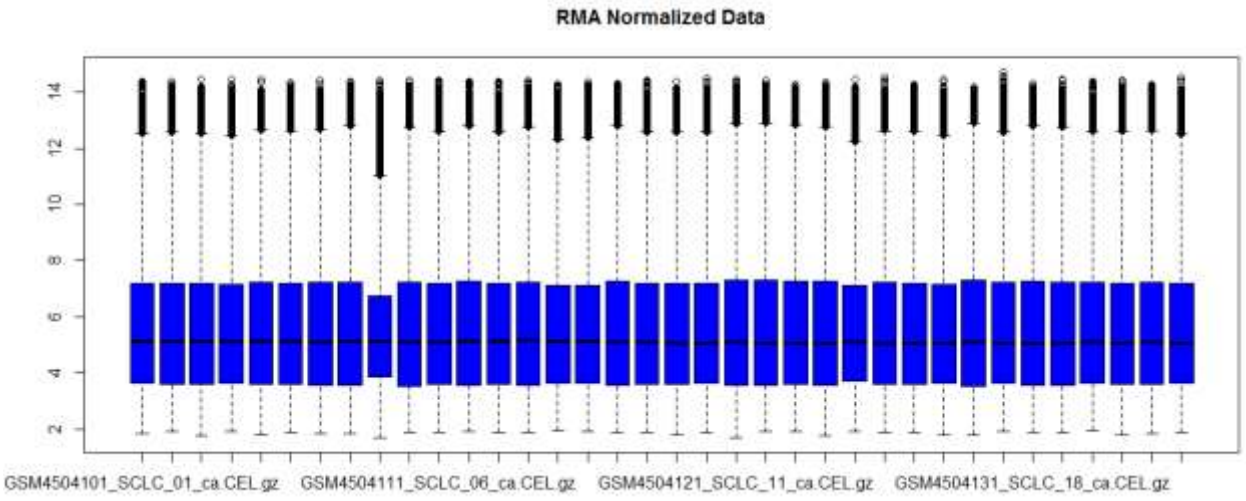


Figure (4): - Box plot before and after normalization

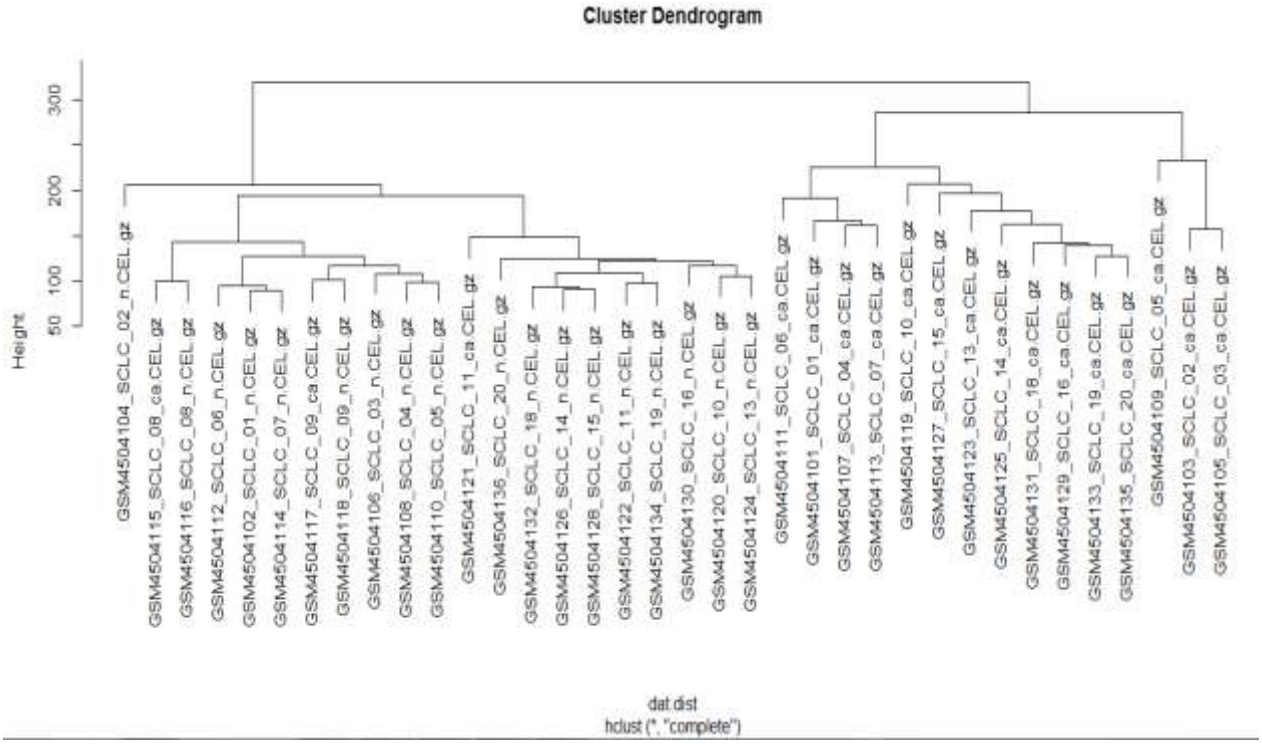


Fig (5): - Clustering of Hierarchical

4- expression filtering genes

When filtering by expression, it's not a good idea to assume that all arrays act the same way. The gene may be expressed on some arrays but may not appear to be expressed at all or at a very low level on others for some reason. As a result, the filter must account for this potential disparity. This is accomplished by allowing a gene to pass the filter (and hence be included in the dataset) if it is expressed at the specified level in at least a percentage of the

samples. During filtering, the former function employs the absolute samples number, while the latter function relies on the proportion. In this case, we'll apply the percentage approach.

5- Analysis using a linear model

The command lmfit() was used for the analysis; this was followed by the eBayes() function. The lmfit() function is looking for the data matrix as well as a design matrix. Once the data is in a matrix format, the analysis can be performed with either the unfiltered or filtered data. The analysis' results are stored in the object fit. The command topTable() can be used to extract results (Figure 6); it produces output for the first design matrix column, i.e. the intercept, by default. The coefficient to report can be altered in the function call using the option Coef, but this coefficient is rarely of interest. In a comparative experiment between two groups, the genes that are statistically significantly differentially expressed between both groups.

1		logFC	AveExpr	t	p.value	adj.P.Val	B
2	202434_s_at	1.475600639	3.986778113	8.829976041	5.44E-10	2.97674244100	05 7.31865235708337
3	208488_s_at	1.277745843	4.674263933	7.249874143	3.60E-08	0.000983244	24 5.1445448632655
4	1569054_at	1.806751073	3.13415083	6.025552688604	.11507604369587	.020322260896	0.167204144
5	217373_x_at	1.144299637	5.239978236	5.756820232	2.40E-06	0.032871826	2.702081175
6	207356_at 0	0.932443544	5.290191921	5.287653921	9.26E-06	0.08725324	1.867895193
7	225282_at 1	.52261147298951 8	.7572121395407 5	2470245379253	04045868483204e	0872532403270	7.94497E+13
8	226368_at 1	.37706545815861 7	.16882737757094	.2223174298159	.11709681260087	.087253240327(	0.749783739
9	1565752_at	1.078748667	5.594398844	5.119431576	1.50E-05	0.099003141	1.562966081
10	217202_s_at	1.7679356	9.477309062	5.0910057	1.63E-05	0.099003141	1.511182899
11	1565754_x_a	t 1.1474797389696	4 5.170585462660	8 4.9962519262	7 2.139834731074	5 0.1146100056	0.338083109
12	204560_at 2	.82359703085135 5	.92176062639506	.9702456272906	30582544537923e	1146100056601	44967463704
13	216233_at 2	.61137909361059 4	.23293021510357	.9221608883372	.64725026250341	.120615340085	2.02246E+13
14	200648_s_at	1.57296815	10.056649	4.848333378	3.27E-05	0.137608272888	0.066509906
15	232725_s_at	1.147898331	3.243090729	4.756220476	4.26E-05	0.158914594075	0.896682916
16	217260_x_at	0.849903426	5 5.267427257360	3 4.7253514244	6 4.653982129406	5 0.1589145940	0.839665803
17	235868_at 1	.1222770526534 3.	19394768145704	7253486432330	6540191732103e-	5891459407530	9.66066E+11
18	231860_at -	1.133491874	6.109155761	-4.704428755	4.94E-05	0.158914594075	0.800993306
19	211913_s_at	1.671106463	6.284902958	4.611638352600	.44146828944574	.195659599291	6.29249E+14

Fig (6): - gene expression table result

6- Gene enrichment and gene annotation

Because the GO hierarchy is made up of three different ontologies, the actual test is done in three steps; these include biological process (BP), cellular component (CC), and molecular function (MF). There were 8374 GO BP ids examined (860 with p 0.05), 1806 GO MF ids tested (141 with p 0.05), and 1074 GO CC ids tested (183 with p 0.05) out of 2500 genes. In the instance of the KEGG 204 ids tested, 20 of them have a p 0.05. If the entire chip is being annotated, building the annotation table could take several minutes, if not an hour.

Bioconductor Affymetrix Probe Listing

Probe	Symbol	Description	Chromosome	Chromosome Location	GenBank	Gene	Cytoband	UniGene	PubMed	Gene Ontology
1552266_at	ADAM32	ADAM metallopeptidase domain 32	8	39107528, 13211	NM_145004	203102	8p11.22	Hs.521545	8	metalloendopeptidase activity proteolysis
1552269_at	SPATA17	spermatogenesis associated 17	1	217631323	NM_138796	128153	1q41	Hs.171130 Hs.665237	12	calmodulin binding cytoplasm
1552271_at	PRR22	proline rich 22	19	-5782959	NM_153359	163154	19p13.3	Hs.631838	3	
1552272_a_at	PRR22	proline rich 22	19	-5782959	NM_153359	163154	19p13.3	Hs.631838	3	
1552274_at	PXK	PX domain containing serine threonine kinase like	3	58332889, 58333305	BC014479	54899	3p14.3	Hs.190544	26	plasma membrane plasma membrane plasma membrane actin binding cytosol cytoplasm cytoplasm microtubule organizing center inflammatory response negative regulation of ATPase activity phosphatidylinositol binding regulation of membrane potential negative regulation of ion transport modulation of chemical synaptic transmission

Acknowledgement

We would like to thanks Abdulrawoof, Hemanshi, Surbala and Mohammed Isam, for their helps.

References

1- B Alex Merrick, Robert E London, Pierre R Bushel, Sherry F Grissom, Richard S Paules, "Platforms for biomarker analysis using high-throughput approaches in genomics, transcriptomics, proteomics, metabolomics, and bioinformatics", IARC Sci Publ 2011;(163):121-42.

2- Andrew J. Yee, Sridhar Ramaswamy, "DNA Microarrays in Biological Discovery and Patient Care" Essentials of Genomic and Personalized Medicine journal, 2010

3- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D.,Winkler, J., Lockhart, D.

1. J., Morris, M. S., Fodor, S. P. (1996) Accessing genetic information with high-density DNA
2. arrays. Science 274, 610-614.

4- Claire L. Wilson and Crispin J. Miller (2005) Simpleaffy: a BioConductor package for

3. Affymetrix Quality Control and data analysis. Bioinformatics 2005 21(18):3683-3685

5- Bolstad BM (2007) affPLM: Model based QC assessment of Affymetrix GeneChips.

6- S. Mase, T. Kamakura, M. Jimbo, and K. Kanefuji. *Introduction to Data Science for*

4. *engineers- Data analysis using free statistical software R (in Japanese)*. Suuri-Kogaku-sha,
5. Tokyo, April 2004. ISBN 4901683128.
7- Brian Everitt and Torsten Hothorn. *A Handbook of Statistical Analyses Using R*. Chapman &
6. Hall/CRC, Boca Raton, FL, 2006. ISBN 1-584-88539-4.
8- Dianne Cook and Deborah F. Swayne. *Interactive and Dynamic Graphics for Data Analysis*.
7. Springer, New York, 2007. ISBN 978-0-387-71761-6.
9- John Maindonald and John Braun. *Data Analysis and Graphics Using R*. Cambridge
8. University Press, Cambridge, 2nd edition, 2007. ISBN 978-0-521-86116-8.
10- Phil Spector. *Data Manipulation with R*. Springer, New York, 2008. ISBN 978-0-387-74730-
9. 9.
11- John M. Quick. *The Statistical Analysis with R Beginners Guide*. Packt Publishing, 2010. ISBN
10. 1849512086.
12- Didkowska, J.; Wojciechowska, U.; Ma´ nczuk, M.; Łobaszewski, J. Lung cancer epidemiology: Contemporary and future challenges worldwide. Ann. Transl. Med. **2016**, 4, 150.
13- Rahal, Z.; El Nemr, S.; Sinjab, A.; Chami, H.; Tfayli, A.; Kadara, H. Smoking and Lung Cancer: AGeo-Regional Perspective. Front. Oncol. **2017**, 7, 194.
14- Hirsch, F.R.; Scagliotti, G.V.; Mulshine, J.L.; Kwon, R.; Curran, W.J.; Wu, Y.L.; Paz-Ares, L. Lung Cancer: Current Therapies and New Targeted Treatments. Lancet **2017**, 389, 299–311.
15- Xi, K.X.; Zhang, X.W.; Yu, X.Y.;Wang,W.D.; Xi, K.X.; Chen, Y.Q.;Wen, Y.S.; Zhang, L.J. The role of plasma miRNAs in the diagnosis of pulmonary nodules. J. Thorac. Dis. **2018**, 10, 4032–4041.
16- Lu, S.; Kong, H.; Hou, Y.; Ge, D.; Huang,W.; Ou, J.; Yang, D.; Zhang, L.;Wu, G.; Song, Y.; et al. Two Plasma microRNA Panels For Diagnosis and Subtype Discrimination of Lung Cancer. Lung Cancer. **2018**, 123, 44–51.
17- Karupusamy, S., Mustafa, M. A., Jos, B. M., Dahiya, P., Bhardwaj, R., Kanani, P., & Kumar, A. (2023). Torque control-based induction motor speed control using Anticipating Power Impulse Technique. The International Journal of Advanced Manufacturing Technology, 1-9.
18- Govindarajan, S., Mustafa, M. A., Kiyosov, S., Duong, N. D., Raju, M. N., & Gola, K. K. (2023). An optimization based feature extraction and machine learning techniques for named entity identification. Optik, 272, 170348.

19- Sudha, I., Mustafa, M. A., Suguna, R., Karupusamy, S., Ammisetty, V., Shavkatovich, S. N., ... & Kanani, P. (2023). Pulse jamming attack detection using swarm intelligence in wireless sensor networks. Optik, 272, 170251.
20- Hassan, J. A., & Rasheed, M. K. (2022, November). Synthesis and characterization of some benzimidazole derivatives from 4-methyl ortho-phenylene diamine and evaluating their effectiveness against bacteria and fungi. In AIP Conference Proceedings (Vol. 2394, No. 1). AIP Publishing.
21- Nijris, O. N., Khaleel, Z. I., Hamady, S. Y., & Mustafa, M. A. (2020). The effectiveness of Aqueous Extract of Grape Seeds Vitis vinifera as an antibiotic for some microorganisms and its Protective Role Histology for Liver, Kidney in Mice. Indian Journal of Forensic Medicine & Toxicology, 14(2), 1838-1845.
22- Mustafa, H. A., Majid, H. H., Abdulqader, A. T., Mustafa, M. A., & Salih, A. A. (2019). Study On Some Physiological, Biochemical And Hormonal Parameters Of Seminal Fluid Of Infertile Men. Biochem. Cell. Arch, 19(Supplement 1), 1943-1947.
23- Fadhil, K. B., Majeed, M. A. A., & Mustafa, M. A. (2019). Electronic study of fresh enzyme complexes of antifungal drugs-P450 and Aspergillus kojic acid biosynthesis. W: w saccharose flavus: fructose as a substratum. Annals of Tropical Medicine and Health, 22, 65-72.
24- Abdulazeez, M., Hussein, A. A., Hamdi, A. Q., & Mustafa, M. A. (2020). Estimate the Complications That Resulting from Delayed Management of Dental Trauma in Tikrit City. Journal of Cardiovascular Disease Research, 11(2), 80-82.
25- Hama Hasan, T. A., Erzaiq, Z. S., Khalaf, T. M., & Mustafa, M. A. (2020). Effect of Equisetum Arvense Phenolic Extract in Treatment of Entamoeba Histolytica Infection. Systematic Reviews in Pharmacy, 11(11).
26- Hama Hasan, T. A., Erzaiq, Z. S., Khalaf, T. M., & Mustafa, M. A. (2020). Effect of Equisetum Arvense Phenolic Extract in Treatment of Entamoeba Histolytica Infection. Systematic Reviews in Pharmacy, 11(11).
27- Nijris, O. N., Khaleel, Z. I., Hamady, S. Y., & Mustafa, M. A. (2020). The effectiveness of Aqueous Extract of Grape Seeds Vitis vinifera as an antibiotic for some microorganisms and its Protective Role Histology for Liver, Kidney in Mice. Indian Journal of Forensic Medicine & Toxicology, 14(2), 1838-1845.
28- Ali, A., Jassim, A.F., Muhsin, S.N., & Mustafa, M.A. (2020). Study of Lycium Shawii Phenolic Compounds in Treatment of Hyperlipidemia. Journal of cardiovascular disease research, 11, 196-199.
29- Ibrahim, H. M., Jumaah, L. F., Khalaf, S. A., & Mustafa, M. A. (2021). KNOWLEDGE AND PRACTICE OF BREASTFEEDING AND WEANING IN MOTHERS LIVES SAMARRA CITY, IRAQ. Biochemical & Cellular Archives, 21.

Cite this article as: **Arshed H.Yaseen (2023).**
Lung cancer data analysis for finding gene expression
African Journal of Biological Sciences. 5(3), 119-130. doi: 10.33472/AFJBS.5.3.2023.119-130