Differential Expression Analysis of Microarray Gene Data for Cancer Detection

Shashank K S[1] Dr Mamatha H R[2]
[1]M. Tech Student (Software Engineering) [2]Professor
[1,2]Department of Information Science and Engineering
[1,2]PES Institute of Technology Bangalore-560085

Abstract— The pre-processed Microarray Gene datasets are utilized for differential quality investigation. Limma bundles are utilized to foresee differential quality expression information emerging from microarray RNA tests. We utilize datasets of control and tumor tests of differential levels, the change between two specimens are recognized through qualities which are up-directed (expanded in expression) or down-controlled (diminished in expression). Grouping of tests with comparative expression designs crosswise over qualities were done. Every specimen gathering will contain various duplicates. The gathering expression level for a test will be compressed as the mean of the expression levels in the gathering reproduces. In this manner, differential expression problems are a comparison of means.

Key words: microarray RNA tests, Microarray Gene datasets, clustering

I. INTRODUCTION

Cancer is an ailment or an unusual state of the cells in human body. There are different sorts of cell in the body to perform particular functions, which in-turn promotes distinctive sorts of disease which influence different sorts of cell bringing on a variation in their actions performed. Yet, cancer disease cells reproduce wild regardless of their cell sort. Some tumor sort have a larger number of genuine mortality impacts than others, some are more effortlessly treated than others (especially if analyzed at an early stage), and a few sorts have a superior standpoint (anticipation/forecast) than other disease sorts. In this way, cancer disease by and large is one condition as well as has numerous related stages. For every situation of disease analyzed it is essential to know precisely what sort of growth has added to, what's the tumor size, is it amiable or dangerous, and how well it more often than not reacts to treatment with a specific end goal to give an effective cure.

A. Pre-Processing

The chosen datasets of diverse test sets were pre-processed utilizing Li and Wong (2001) [1] (quartile) standardization calculation. The pre-processing of crude (raw) test (probe) level information of raw signal intensities for every test sets of all CEL records were standardized (normalized) to correct background intensity calculations. Various algorithms for example, MAS5, RMA [2][3][4] were used. The foreground adjustment of crude intensity information calculations, used algorithms such as consistent, quintiles and invariant set. After the pre-preprocessing of standardized information the ideal(perfect) match correction utilizing pmonly and MAS5 calculation [2][3][4] was carried out. The general pre-processing of all dataset outline were anticipated utilizing MAS, Li and Wong (2001) [1] and median polish algorithms [5]. All the pre-processing of analysis is done utilizing BioConductor Packages. Utilizing Affy and affycoretools bundle to correct background and test intensities with RMA, MAS5 calculations of background amends, standardizes and abridges the test levels. Then again, the signal force for MM test can frequently be bigger than PM test inferring that MM test is recognizing genuine sign and additionally background (foundation) signal

B. Related Work

In the present year 9.1 million women's is affecting with breast cancer in worldwide. In addition 232,670 women's is diagnosed with in a year. 30% of women population is affected due to genetic abnormalities such as mutation of BRCA1 and BRCA2 genes [6]. In addition there are some other oncogenic genes such as k-RAS, p53, PTEN, NBS1 causing breast cancer [7]. The colon cancer is also a leading cancer types that frequently affects other tissues such as lungs, breast and prostate tissues. The genetic alterations of k-RAS, APC, P53, β-catenin, GSK-3β that mainly affect WNT- β-catenin signaling pathways that also affects breast and ovarian cancer [8] [9]. The epithelial ovarian cancer is also dangerous cancer type in women, the mutation of p53, BCL-XL, EGFR, MDM2, MCI-2, NOXA and others is mainly involved in ovarian cancer [10][11][12]. A number of genetic marker has been proposed to identify cancer such as BRCA1, BRCA2 of breast cancer, APC, GSK-3β of colon cancer and CA125 of ovarian cancer. In addition there are more number of serum markers that helps to clinically diagnostics through breast, colon and ovarian cancer. However, its effectiveness of detecting more genetic markers that widely considered to use more advanced techniques such as DNA microarray technology to identify genetic profiling of cancer by allowing thousands of genes that significantly associated with cancer types [13][14].

II. DIFFERENTIAL GENE ANALYSIS

Pre-processing of datasets will bring about CEL documents that can be further utilized for differential quality expression. Limma bundles of R device are utilized to foresee differential quality expression information emerging from microarray RNA tests. For datasets of control and growth tests of differential levels, the change between two examples are recognized utilizing qualities which are up-directed (expanded in expression) or down-managed (diminished in expression). The grouping of qualities that takes after expression designs over an arrangement of tests, or bunching specimens with comparative expression designs crosswise over qualities are additionally a piece of the differential quality examination. Every specimen gathering will contain various imitates. The gathering expression level for a test will be condensed as the mean of the expression levels in the gathering reproduces. In this manner,

differential expression issues are a correlation of means. At the point when there are two specimen amasses this is a test or something to that affect. The bunching of up directed and down managed qualities were anticipated with grouping analysis.

Hierarchical clustering of differentially expressed genes with respect to probable expression of B values with correlation coefficient of control-cancer datasets. The relationship among objects are represented by a tree whose branch lengths id different with differential gene expression. The differential communicated qualities were explained utilizing GO.db bundle.

The Annotation of HGU 133plus 2 bundle of GO annotation serves to comprehend the qualities included in differential communicated qualities alongside natural procedure, sub-atomic capacity or cell parts of qualities with precise characterization.

III. Materials and Methods

The differential expression analysis of three different cancer diseases such as breast, colon and ovarian cancer datasets were retrieved from GEO database. The Datasets of breast cancer (GEO ID: GSE30543) of 6 samples with SUM149 control siRNAs and siRNA targeting TIG1 replicates [15].The colon cancer dataset (GSE34299) of 4 samples has HT29 parental cell lines and HT29RC PLX4720 resistant cell lines grown in increasing concentration of the drug to develop acquired resistance [16]. The ovarian cancer dataset (GSE35972) of 6 samples has untreatedTOV112D cells and NSC319726 treated with different biological replicates [17].

IV. Experimental Design

The pre-processed datasets are utilized for differential quality expression examination utilizing Limma bundle. We pre-prepared 54675 qualities to discover differential quality expression. The pair wise examination of diverse datasets taking into account log2 fold changes, standard slips, t-insights and p-values. The rundown of the outcomes were recorded in table.We utilized fundamental measurements for importance investigation the directed t-measurement, which is processed for every test and for every differentiation. This has the same translation as a conventional t-measurement with the exception of that the standard lapses have been directed crosswise over qualities, i.e., contracted towards a typical worth, utilizing a basic Bayesian model. This has the impact of getting data from the troupe of qualities to help with deduction about every individual quality. Directed t-measurements lead to p-values in the same way that conventional t-insights do aside from that the degrees of flexibility are expanded, mirroring the more noteworthy unwavering quality connected with the smoothed standard slips. The viability of the directed t methodology has been exhibited on test information sets for which the differential expression status of every test is known.

Various synopsis measurements are displayed by topTable() for the top qualities and the chose contrast. The differential articulation of both up directed and down controlled gens were recorded in view of logFC segment gives the estimation of the difference. Typically this speaks to a log2-fold change between two or more trial conditions albeit now and then it speaks to a log2-expression level. The AveExpr segment gives the normal log2-expression level for that quality over every one of the exhibits and diverts in the investigation. Section t is the directed t-measurement. Section P.Value is the related p-quality and adj.P.Value is the p-worth balanced for numerous testing. The most prominent type of conformity is "BH" which is Benjamini and Hochberg's system to control the false disclosure rate. The balanced qualities are regularly called q-values if the aim is to control or assessment the false disclosure rate. The importance of "BH" q-qualities is as per the following. In the event that all qualities with q-esteem underneath a limit, say 0.05, are chosen as differentially communicated, then the normal extent of false revelations in the chose gathering is controlled to be not exactly the edge esteem, for this situation 5%. This methodology is equal to the system of Benjamini and Hochberg despite the fact that the first paper did not plan the strategy regarding balanced p-values [18].

V. Experimental Results

A. Differential Gene Expression Analysis On Breast Cancer

The analysis of breast cancer dataset contains 6 samples such as 3 SUM149 cells transfected with control sample of siRNA and SUM149 cells transfected with siRNA targeting of tarzarotene-induced gene 1 (TIG1). All these 6 samples is annotated with hgu133plus2 contains 54675 genes, using normalization methods to filter the genes that significantly associated with p-values, we have filter the 54675 genes of which 12788 genes that has significantly associated with gene expression. 1220 genes has upregulated and 11568 genes is downregulated that differentially expressed in breast cancer

ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
210176_at	-0.518098	2.31404253	-14.128827	2.75E-09	0.0001190	9.55460
220317_at	0.422659	2.375689545	12.948213	8.01E-09	0.0001190	8.91178
203854_at	-0.526925	2.538701527	-12.860364	8.70E-09	0.0001190	8.86005
202376_at	-0.340333	2.8862006	-12.697768	1.02E-08	0.0001190	8.76277
232318_s_at	-0.502046	2.436144194	-12.625586	1.09E-08	0.0001190	8.71893
211343_s_at	0.573454	2.308990431	12.37709828	1.38E-08	0.0001261	8.56493
225801_at	-0.534082	2.357775707	-11.419254	3.64E-08	0.0002846	7.92332

220232_at	-0.323877	2.758999409	-10.043355	1.66E-07	0.0011331	6.85031
218806_s_at	-0.279414	2.660891323	-9.7767399	2.27E-07	0.0013768	6.61903
229125_at	-0.308686	2.505980042	-9.5716889	2.90E-07	0.0015830	6.43553
201893_x_at	-0.254423	3.335987394	-9.43194735	3.43E-07	0.0015831	6.30760
205363_at	-0.450918	2.355673702	-9.3771957	3.67E-07	0.0015831	6.25683
241762_at	-0.826396	2.375262471	-9.3556430	3.76E-07	0.0015831	6.23674
219934_s_at	-0.383661	2.338478608	-9.2119730	4.49E-07	0.0017548	6.10137
214701_s_at	-0.300928	2.935614599	-9.1242836	5.01E-07	0.0018267	6.01746
211813_x_at	-0.262725	3.310657648	-9.0626261	5.41E-07	0.00185	5.95789
219276_x_at	0.283229	3.258745992	8.8151891	7.41E-07	0.0022979	5.71383
214961_at	-0.257423	2.157925543	-8.7989439	7.57E-07	0.0022979	5.69753
208747_s_at	-0.373868	2.782926507	-8.7500092	8.06E-07	0.00231832	5.64821
204140_at	-0.212168	3.08483983	-8.6868001	8.74E-07	0.0023897	5.58402

Table 5.1: Upregulated genes predicted in breast cancer data

B. Differential Gene Expression on Colon Cancer Data

Our analysis was done usig 4 datasets of which 2 samples are HT29 parental cell lines and another two samples are HT29RC PLX4720 resistant cell lines. We have found 268 upregulated genes is consistently expressed in HT29RC cell lines. There are 1268 down regulated genes also associated with both cell lines and control tissues

Genes	logFC	Fold Change	AveExpr	t	P.Value	adj.P.Val	B
GPX2	-10.63	-1590.5	5.0346	-86.99	1.42E-21	2.52E-17	37.2784
KRT81	10.15	1140.91	7.8215	83.712	2.51E-21	2.52E-17	36.9216
LYZ	-10.86	-1860.89	5.4100	-80.013	4.91E-21	3.29E-17	36.4886
KRT6C	11.47	2847.97	6.15143	77.773	7.49E-21	3.77E-17	36.208719
GIPC2	-6.77	-109.30	3.88231	-75.58	1.14E-20	4.60E-17	35.9216
BNC1	8.414	341.150	5.75057	66.732	7.28E-20	2.12E-16	34.6001
KRT6A	10.95	1990.34	6.34167	66.672	7.38E-20	2.12E-16	34.590146
TOX3	-9.66	-810.56	4.74252	-65.28	1.01E-19	2.54E-16	34.3557
NA	-8.98	-506.22	7.66447	-54.88	1.32E-18	2.69E-15	32.3259
NA	-9.55	-753.25	8.23008	-54.85	1.34E-18	2.69E-15	32.31773
TSPAN8	-11.23	-2410.43	5.32834	-54.33	1.54E-18	2.81E-15	32.20090
NA	-8.74	-429.45	7.50284	-53.00	2.22E-18	3.72E-15	31.89581
LGALS3BP	-6.64	-100.21	4.70832	-51.67	3.23E-18	5.00E-15	31.58052
HTRA1	9.61	786.58	5.87828	50.912	4.03E-18	5.79E-15	31.39287
ACSL5	-8.21	-296.51	5.08635	-50.49	4.55E-18	6.10E-15	31.28981
RAB34	8.62	394.79	8.76300	48.514	8.23E-18	1.03E-14	30.77931
CYBRD1	6.88	118.19	7.40537	46.544	1.52E-17	1.80E-14	30.24366
SLC40A1	-8.69	-415.81	7.1428	-44.59	2.86E-17	3.20E-14	29.68375
TSPYL5	5.71	52.46	4.6624	42.629	5.58E-17	5.72E-14	29.08408
LGALS4	-9.79	-889.68	4.7316	-42.57	5.69E-17	5.72E-14	29.06581

Table 5.2: Upregulated genes predicted in colon cancer

C. Differential Gene Expression on Ovarian Cancer Data

The ovarian cancer studies has 6 samples of which 3 samples is p53 targeted TOV112D cells untreated and other 3 samples P53 targeted NSC319726 cells data. There are total of 54675 genes of which 1566 genes is significantly associated differential expression in ovarian cancer. There are 810 genes is upregulated within ovarian tissues of which p53 associated with cancer and control tissues, 756 genes down regulated based on significant test.

ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
209173_at	-0.518098	2.31404253	-14.128827	2.75E-09	0.0001190	11.5546
206378_at	0.422659	2.375689545	12.948213	8.01E-09	0.0001190	10.91178

240304_s_at	-0.526925	2.538701527	-12.860364	8.70E-09	0.0001190	10.86005
209602_s_at	-0.340333	2.8862006	-12.697768	1.02E-08	0.0001190	10.76277
228241_at	-0.502046	2.436144194	-12.625586	1.09E-08	0.0001190	10.71893
205048_s_at	0.573454	2.308990431	12.37709828	1.38E-08	0.0001261	10.56493
227662_at	-0.534082	2.357775707	-11.419254	3.64E-08	0.0002846	9.92332
228915_at	-0.323877	2.758999409	-10.043355	1.66E-07	0.0011331	8.85031
214451_at	-0.279414	2.660891323	-9.7767399	2.27E-07	0.0013768	8.61903
237339_at	-0.308686	2.505980042	-9.5716889	2.90E-07	0.0015830	8.43553
219580_s_at	-0.254423	3.335987394	-9.43194735	3.43E-07	0.0015831	8.3076
205239_at	-0.450918	2.355673702	-9.3771957	3.67E-07	0.0015831	8.25683
237086_at	-0.826396	2.375262471	-9.3556430	3.76E-07	0.0015831	8.23674
204667_at	-0.383661	2.338478608	-9.2119730	4.49E-07	0.0017548	8.10137
223864_at	-0.300928	2.935614599	-9.1242836	5.01E-07	0.0018267	8.01746
214774_x_at	-0.262725	3.310657648	-9.0626261	5.41E-07	0.00185	7.95789
216623_x_at	0.283229	3.258745992	8.8151891	7.41E-07	0.0022979	7.71383
226597_at	-0.257423	2.157925543	-8.7989439	7.57E-07	0.0022979	7.69753
238778_at	-0.373868	2.782926507	-8.7500092	8.06E-07	0.00231832	7.64821
238017_at	-0.212168	3.08483983	-8.6868001	8.74E-07	0.0023897	7.58402

Table 5.3: Upregulated genes predicted in ovarian cancer data

VI. CONCLUSIONS

The datasets of control and cancer samples of differential levels were successfully analysed using the differential analysis technique. The change between two samples is identified through genes which are up-regulated (increased in expression) or down-regulated (decreased in expression). Clustering of those samples with similar expression patterns across genes were carried out. The outcome this result further helps in analysing the comparative analysis of differentially expressed genes using various classifiers and providing the gene network along with functional annotation and enrichment analysis.

REFERENCES

[1] Cheng Li and Wing Hung Wong Departments of Statistics and Human Genetics, University of California, Los Angeles, CA 90095“ Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection” Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved October 30, 2000 (received for review August 21, 2000)

[2] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P.Speed. “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias” . Bioinformatics, 19(2 :185-193,Jan 2003.

[3] Rafael A. Irizarry, Bridget Hobbs, Francois Collin,Yasmin D. Beazer-Barclay, Kristen J. Antonellis,Uwe Scherf, and Terence P. Speed. “Exploration,normalization, and summaries of highdensity oligonucleotide array probe level data”. Biostatistics,2003b.

[4] Rafael A. Irizarry, Benjamin Bolstad, Francois Collin, Leslie Cope, Bridget Hobbs and Terence Speed “Summaries of Affymetrix GeneChip probe level data” Nucleic Acids Research 31(4)2003

[5] Federico M Giorgi, Anthony M Bolger, Marc Lohse and Bjoern Usadel “Algorithm-driven Artifacts in median polish summarization of Microarray data” © 2010 Giorgi et al; licensee BioMed Central Ltd.

[6] Dumitrescu RG, Cotarla I. Understanding breast cancer risk--where do we stand in 2005? J Cell Mol Med. 2005;9:208–21

[7] Honrado E, Osorio A, Palacios J, Benitez J. Pathology and gene expression of hereditary breast tumors associated with BRCA1, BRCA2 and CHEK2 gene mutations. Oncogene. 2006;25:5837–45

[8] Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM and Bos JL. (1988). N. Engl. J. Med., 319, 525–532.

[9] Fearon ER and Vogelstein B. (1990). Cell, 61, 759–767.

[10] Baekelandt M et al. (1999) Clinical significance of apoptosis-related factors p53, Mdm2, and Bcl-2 in advanced ovarian cancer. J Clin Oncol 17: 2061

[11] Kupryjanczyk J et al. (2003) Evaluation of clinical significance of TP53, BCL-2, BAX and MEK1 expression in 229 ovarian carcinomas treated with platinum-based regimen. Br J Cancer 88: 848–854

[12] Nielsen JS et al. (2004) Prognostic significance of p53, Her-2, and EGFR overexpression in borderline and epithelial ovarian cancer. Int J Gynecol Cancer 14: 1086–1096
[13] Golub TR (2001) Genome-wide views of cancer. N Engl J Med 344: 601–
[14] Elvidge G (2006) Microarray expression technology: from start to finish. Pharmacogenomics 7: 123–134.
[15] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30543
[16] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34299
[17] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35972
[18] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289–300.

