International Journal of Statistics in Medical Research, 2017, 6, 1-9 1

E-ISSN: 1929-6029/17

Evaluation of Methods for Gene Selection in Melanoma Cell Lines

Linda Chaba1, John Odhiambo1 and Bernard Omolo2,*

1Strathmore Institute of Mathematical Sciences, Strathmore University, Ole Sangale Road, Nairobi, Kenya 2Division of Mathematics & Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, South Carolina, USA Abstract: A major objective in microarray experiments is to identify a panel of genes that are associated with a disease outcome or trait. Many statistical methods have been proposed for gene selection within the last fifteen years. While the comparison of some of these methods has been done, most of them concentrated on finding gene signatures based on two groups. This study evaluates four gene selection methods when the outcome of interested is continuous in nature. We provide a comparative review of four methods: the Statistical Analysis of Microarrays (SAM), the Linear Models for Microarray Analysis (LIMMA), the Lassoed Principal Components (LPC), and the Quantitative Trait Analysis (QTA). Comparison is based on the power to identify differentially expressed genes, the predictive ability of the genelists for a continuous outcome (G2 checkpoint function), and the prognostic properties of the genelists for distant metastasis-free survival. A simulated dataset and a publicly available melanoma cell lines dataset are used for simulations and validation, respectively. A primary melanoma dataset is used for assessment of prognosis. No common genes were found among the genelists from the four methods. While the SAM was generally the best in terms of power, the QTA genelist performed the best in the prediction of the G2 checkpoint function. Identification of genelists depends on the choice of the gene selection method. The QTA method would be preferred over the other approaches in predicting a quantitative outcome in melanoma research. We recommend the development of more robust statistical methods for differential gene expression analysis. Keywords: Differential gene expression, Melanoma cell lines, Prediction, Power, Quantitative trait.

1. BACKGROUND

Microarray technology has revolutionized genomic studies by enabling the study of differential expression of thousands of genes simultaneously. In the recent past, a number of statistical methods have been developed for class comparison and prediction, based on the gene expression profiling of tumors, cell-types, etc. One of the early methods developed was the foldchange method. This method did not account for statistical variation across the samples and suffered from bias if the data were not properly normalized [1].

A number of articles have provided a survey of different statistical methods for finding differentially expressed genes (DEGs) [1-8]. Jeffery et al. [4] compared the efficiency of 10 feature selection methods and applied the methods to 9 different binary (two class) microarray datasets. Bair [7] discussed a number of statistical methods, including fold change, methods

based on the t-test and Bayesian methods that can be used to find differentially expressed genes. However, he did not compare their performance on any dataset. Bandyopadhyay et al. [8] reported a comprehensive survey of different parametric and nonparametric testing methodologies used for finding

*Address correspondence to this author at the Division of Mathematics & Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, South Carolina, USA; Tel: +1 864-503-5362; Fax: +1 864-5035930; E-mail: bomolo@uscupstate.edu

DEGs from microarray datasets. Like in most of the studies, they did not exhaust all the available methods for finding DEGs.

Despite all the surveys mentioned above, there is no unanimous agreement on any particular gene selection method as the standard. A review and comparison of the statistical methods may provide bioinformaticians and other biomedical researchers with a useful guide for choosing the right method for the right data in differential gene expression analysis. Furthermore, even though work has been done on the development of methods for the differential analysis of gene expression data measured in two conditions, open research questions still exist regarding the analysis of gene expression data in which the training signal is a continuous variable.

This study reports a comparative review of four methods (SAM, LIMMA, LPC and QTA) and their performance in identifying genes that are associated with a continuous outcome from the systems biology of melanoma, using a larger number of melanoma celllines than reported in [9]. While the comparison of some of these methods has been done, most of them concentrated on finding gene signatures based on two groups. A comparison of the LPC method with other methods is conspicuously missing in almost all the surveys presented in the literature. Furthermore, the available studies do not assess the biological and clinical significance of the genes generated by theses

methods. Our study attempts to fill this gap in the literature. The comparison is based on the size and the statistical assessment of the predictive and the prognostic properties of the genelists produced by these methods. 2. OVERVIEW OF METHODS 2.1. Statistical Analysis of Microarrays (SAM)

The SAM method was originally developed to identify genes that are differentially expressed by incorporating a set of gene-specific t-tests [10]. Although Tusher et al. [10] analyzed a two-state experiment (with a dichotomous covariate or response), the SAM procedure can be applied to studies with continuous responses as well. The SAM method identifies DEGs by use of gene-specific moderated t-tests on the basis of the regression coefficient relative to the standard deviation of repeated expression measurements for that gene. SAM employs the false discovery rate (FDR) control for the multiple testing problem and estimates the FDR through the

permutation of values of the response variable and the moderated t-tests. The SAM method is implemented in the R package called samr. 2.2. Linear Models for Microarray Analysis (LIMMA)

LIMMA is an R package that integrates a number of statistical methods to effectively analyse large gene expression data [11]. LIMMA fits a linear model for each gene, given a series of arrays, and uses the Empirical Bayes (EB) method [12] to estimate the posterior variance for each gene [13, 14]. The use of the EB method allows combination of information across genes thus improving variance estimation. To assess the significance of each gene, the moderated tstatistics and their associated p-values are generally used [14]. limma calculates the Bayesian log -odds of differential expression for each gene. The higher the value of the log-odds, the more significant the result. The family-wise error rate (FWER) and the FDR are used in multiple testing adjustment. The LIMMA method is implemented in the R package called limma. 2.3. Lassoed Principal Components (LPC)

The lassoed principal components (LPC) method involves using existing gene-specific scores (T) to calculate scores that provide a more accurate ranking of genes as differentially expressed [15]. Some of the gene-specific scores can be calculated using LIMMA [11], SAM [10] and standardized regression methods,

among others existing methods. LPC identifies significant genes based on the values of the FDRs. It estimates its FDR based on an adjustment of the FDR of the T [15]. The LPC method does not assume that genes are independent but rather takes into account that they work in pathways. The LPC method is similar to the LIMMA method in that they both combine information, or borrow strength, across genes. They do not also do permutation-based inference. The LPC algorithm is implemented in both the R package called lpc [15] and the BRB-ArrayTools software [16]. 2.4. Quantitative Trait Analysis (QTA)

This approach finds genes that are significantly correlated with a quantitative outcome such as age. It uses the correlation coefficient as a measure of dependence to compute p-values. The two most commonly used correlation coefficients are the Pearson's correlation coefficient and the Spearman's (rank) correlation coefficient.

There are two ways of controlling the number of false discoveries in the QTA approach. The first one is based on the p-values computed from the parametric t- or F-tests. Here, a stringent p-value threshold (say $p<0.001$), is used in controlling the number of false discoveries. The second approach uses multivariate permutation tests [17]. The QTA method is implemented in the BRB-ArrayTools software [16].

A concise summary of the four statistical methods is provided in Supplementary Table T1. 3. SIMULATED GENE EXPRESSION DATA We conducted a simple simulation study to compare the four methods in terms of power. Let n and G denote the number of samples and genes, respectively. Further, let D denote the number of genes assumed to be truly differentially expressed. Then (G!D) genes are assumed to be non-differentially expressed. The gene expression data matrix, X, is a G!n matrix of log2-ratios. We can write X asX=(X1,X2), where X1 and X2 are D!n and (G!D)"n matrices, respectively. We set D=50, and n=35 and G to be

1000. We generated the (1000!D) genes from the standard normal distribution. To generate the D genes, we used the standard normal distribution in conjunction with the Cholesky decomposition [18] of their correlation matrix as follows: 1. We generate an unstructured correlation matrix !. ! is a (D+1)!(D+1) matrix that has (i,j)th element given by ! i,j :=corr(xi,xj)

Evaluation of Methods for Gene Selection in Melanoma Cell Lines International Journal of Statistics in Medical Research, 2017, Vol. 6, No. 1 3 2. Find the Cholesky factor, A, of ! such that !=AA'. 3. Let zi :N(0,In),i=1,2,...,(D+1). 4. Z=(z1,z2,...,zD+1 !) 5. XD+1=AZ. XD+1 is the gene expression matrix for D genes that are assumed to be differentially expressed or significantly correlated with the covariate y. y can take any of the D+1 row vectors from the matrix XD+1. X1 is therefore a submatrix of XD+1 with dimensions D!n. An R code for the above simulation is available from the authors.

All the four methods were applied to the simulated data. Differentially expressed genes were identified based on the methods' estimated FDR values. A gene was differentially expressed if its estimated FDR was less that a pre-specified value α. Power was calculated as the ratio of the number of correctly identified differentially expressed genes, true positives (TP), to the total number of truly differentially expressed genes, 50 [19]. 4. APPLICATION

The four methods were applied to the melanoma cell lines dataset to identify DEGs. The genelists generated by the four methods were then applied to an independent melanoma dataset for prognostic assessments. Below are the descriptions of the datasets used in the application. 4.1. Data

Melanoma Cell Lines Dataset The gene expression data (raw intensities) consists of 54 cell-lines (35 melanoma cell lines and 19 normal human melanocytes (NHMs)), each with 45,015 probes. This data is publicly available from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE40047. Only the melanoma cell lines were analyzed. The raw dataset was median-normalized and log2, transformed. If multiple probes mapped to the same gene symbol, they were reduced to one per gene symbol by using the most variable probe(set) measured by interquartile range (IQR) across arrays. Filtration and normalization of the gene expression data was implemented using BRB ArrayTools software [16]. A gene was filtered out if less than 20% of its expression data values had at

least 1.5-fold change in either direction from the genes median value. Genes with more than 50 % missing data across all its samples were also filtered out. There were 3,860 genes available for subsequent analysis.

G2 Checkpoint Function Having obtained the gene expression data, we needed to quantify the biological process in melanoma progression. We selected the G2 checkpoint function in this regard. The G2 checkpoint is a position of control in the cell cycle that delays or arrests mitosis when DNA damage by radiation is detected. The G2 checkpoint prevents cells with damaged DNA cell from entering mitosis, thereby providing the opportunity for repair and stopping the proliferation of damaged cells. Figure 1 below shows the four phases of the cell cycle, including the location of the G2 checkpoint as the last checkpoint before mitosis. The G2 checkpoint function scores were obtained from Kaufmann's lab (UNC - Pathology and Lab Medicine) and

had been calculated as ratios of mitotic cells in 1.5 Gy ironizing radiation (IR)-treated cultures in comparison to their sham-treated control (i.e. IR to sham ratio) [20]. It had been shown in Omolo et al. [20] that the G2 gene signature was prognostic for the development of distant metastasis, hence the choice of G2 checkpoint function for this study.

Figure 1: Cell cycle. After completing DNA synthesis and progression through the G2 phase, the cell enters the mitotic phase, where the chromosomes segregate into two daughter cells. Image downloaded from http://www.bristol.k12.ct.us/ page.cfm?p=7093.

Independent Melanoma Dataset An independent data set, consisting of gene expression data from 6307 genes on 58 primary

melanomas with survival outcome, was obtained for assessing prognosis of the gene signatures from the four methods. This dataset has been reported in [21] and will hereafter be referred to as the Winnx dataset. This data is publicly available in the Array Express data repository at the European Bioinformatics Institute (http://www.ebi.ac.uk/arrayexpress/) under the accession numbers: E-TABM-1 IGR_MELANOMA_ STUDY. The primary endpoint for the study was a 4year distant metastasis-free survival (DMFS), which was defined as the time interval between the diagnosis of the primary cutaneous melanoma and a distant metastasis or death from melanoma within 4 years. Patients alive at the date of last follow-up were censored at that date. Patients were also separated into two groups, one group with distant metastasis-free survival of more than 4 years (group M-) and one group with distant metastasis-free survival of 4 years or less (group M+). 4.2. List of DEGs

To find DEGs, we applied different software for different methods. For the LIMMA approach, we used the limma R package. We fixed the degrees of freedom for the design matrix to be 5. For the SAM approach, the samr R package was used. ! was fixed at 0.00, to allow a large list of DEGs to be generated at different estimated FDR values. The number of nearest neighbors to use for imputation of missing features (knn.neighbors) was set at 10 and the number of permutations was fixed at 1000. The QTA method assessed significance of correlation based on the Spearman's correlations and implemented the procedure using the BRB-ArrayTools software. Similarly, the LPC method was implemented by the BRB-ArrayTools software. The number of DEGs were generated at various levels of estimated FDR threshold (0.01, 0.05 ,0.1, 0.2) for all the methods. 4.3. Prediction and Prognosis

We assessed the predictive quality of each of the genelists by its mean squared error (MSE) of prediction of the G2 checkpoint function. For this, linear models containing significant genes were formulated. Since G>>n, the least absolute shrinkage and selection operator (LASSO) algorithm [22] was used to select genes to include in the models. LASSO builds a sequence of models containing upto n genes and index by F, the number of algorithmic steps relative to the model containing n genes (full model). For each F, a cross-validation estimate is obtained using the leave

one-out cross-validation (LOOCV) method. The final model selected corresponds to the F-value with the minimal estimated mean squared error.

We performed a survival risk prediction (SRP) to assess the clinical significance of the genelists using the Winnx dataset. The clinical outcome for this dataset was 4-year distant metastasis-free survival (DMFS) and the objective was to predict a patient's risk (low/high) for developing distant metastasis within 4 years of primary diagnosis. The SRP procedure entails first reducing the number of candidate genes to only the Cox ones, using the supervised principal component (SPC) method of [23]. These Cox genes are then used to compute the prognostic index for each sample. Samples (patients) with a prognostic index above the median are classified as high risk; otherwise, they are low risk. A log-rank test is performed to test if the two survival curves for the low- and the high -risk groups are significantly different, using the original DMFS values. A genelist would be prognostic for DMFS if the log-rank test is significant. The entire SRP procedure was implemented by a tool of the same name in BRB-ArrayTools software [16]. We compared the performance of the genelists produced by the four methods in survival risk prediction for the 58 samples in the Winnx dataset.

In addition, we used the Prediction Analysis of Microarrays (PAM) tool to predict the group membership of the 58 samples. Samples were grouped into two classes: a group with distant metastasis-free survival of more than 4 years (group M-) and a group with distant metastasis-free survival of 4 years or less (group M+). PAM uses the shrunken centroid algorithm developed by [24]. This algorithm builds a number of linear models and selects the model with the least prediction error. A cross-validation estimate is obtained by using leave-one-out cross-validation (LOOCV). The entire model building process is repeated for each leave-one-out training set. The misclassification rate was calculated as the proportion of times the models incorrectly predict the class of the excluded samples. The genelist with the lowest misclassification rate was considered a good list for predicting a sample as belonging to group M+ or M-. 5. RESULTS AND DISCUSSION 5.1. Differentially Expressed Genes

Each of the four methods was applied to the simulated data. The total number of genes that were

correctly identified as differentially expressed, true positives (TP), were recorded at different estimated FDR levels. With the known number of TP, the power was also calculated to aid in comparison. Table 1 shows the number of DEGs and power by different methods at different FDR levels. The LPC method turns out to be the least powerful of all the methods. The SAM and the QTA methods are the most powerful methods in the identification of DEGs. The LIMMA method has moderate power (>0.7) for the FDR thresholds considered, except at the FDR <0.01.

Although the SAM and the QTA methods performed the best with the simulated dataset, we needed to determine how they perform on a real dataset. We applied the methods to the melanoma cell lines dataset (in 4.1). The results are different from the ones obtained using the simulated dataset. We observe that while the QTA method did well with the simulated dataset,

its performance is the worst in the identification of DEGs using the real dataset. In terms of power, the SAM method is still the best followed by the LIMMA method.

The difference in the performance of the QTA method when applied to the simulated and the real datasets could be explained by the fact that the simulated dataset is generated from a standard normal distribution. The QTA method strongly assumes that the gene expression levels (log2-ratios) are normally distributed. Gene expression data may violate this assumption. The LPC method assumes that a large set of genes work together in a pathway to cause an outcome. In cases where this assumption is not met i.e. when only one gene or very few genes cause the outcome, the LPC method loses power in selecting significant genes. This could explain the low performance of the LPC method in both simulated and real datasets. One disadvantage of the LPC method is that it does not rank genes using a metric that is relevant or truly of interest. It rather finds genes that generate high values when standard scores are

projected into a high-variance subspace of the gene expression data [15].

Since different spots on the microarrays are assumed to contain different probes (in the case of cDNA arrays) or different oligos (in the case of highdensity oligonucleotide arrays), the expression of genes are assumed independent on these spots, even though some probes may represent the same gene and have dependent expression profiles. Consequently, not all the methods for selecting DEGs assume that the genes are independent. In particular, the LPC method does not assume that the genes are independent, while the SAM and the QTA methods do assume independence. While the LIMMA approach assumes independence, it works well when the genes are assumed dependent as well [13]. This has been one of the main differences among the four methods.

Before analyzing the validation datasets, the gene expression data were filtered and normalized to eliminate genes that were not sufficiently differentially across the samples and to correct for sample-specific bias (due to experimental artefacts/errors) and render the samples comparable, respectively. After normalization, the resulting expression data was log2transformed so as to achieve a symmetric error distribution. Supplementary Figure F1 shows the error distribution for four randomly selected melanoma cell lines and primary tumors as symmetric and can be regarded as “approximately" normal.

Figure 2 Shows the number of overlapping genes from the four methods. It is very common to find a very low number of overlapping DEGs between multiple methods [4, 25]. 5.2. Prediction and Prognosis

We used the genelists generated by the four methods to build linear predictive models for the G2 checkpoint function, via the LASSO with LOOCV. Table 2 provides a summary of the results. The QTA genelist

Table 1: Number of DEGS Generated by the SAM, LIMMA, LPC and QTA Methods at Different Levels of Estimated FDR (α). The Power of the Methods are Shown (in Parenthesis), Based on the Simulated Dataset

Simulated dataset	Melanoma dataset

α

SAM LIMMA LPC QTA SAM LIMMA LPC QTA .01 49 (0.98) 6 (0.12) 0 (0.00) 50(1.00) 0 8 3 0 .05 52 (1.00) 39 (0.74) 0 (0.00) 51(1.00) 33 16 4 0 .1 53 (1.00) 50 (0.88) 0 (0.00) 54 (1.00) 33 22 7 4 .2 56 (1.00) 57 (0.82) 0 (0.00) 67 (1.00) 173 55 24 56

turns out to be the best in predicting G2 followed by the SAM genelist, then the LIMMA genelist. In order to get additional insight into the performance of the four methods, the four genelists were combined to get 52 unique genes. This combined genelist yielded an R2 of 0.506. A combination of all the genelists had a much better performance than most of the genelists generated by the individual methods.

Gene expression data for the four genelists were extracted from the Winnx dataset for performing survival risk prediction. The difference between the survival curves for the low- and high -risk groups is significant for the SAM genelist (log-rank ! 2 =.5,p=0.019), the LPC genelist (log-rank ! 2 =5.7,p=0.0166) and the QTA genelist (log-rank ! 2 =4.8,p=0.0374) but not for the LIMMA genelist (log-rank ! 2 =0.1,p=0.791). Results are shown in Figure 3.

We further subjected the combined genelist to a survival risk prediction analysis using the Winnix dataset. This genelist provides a good prediction of the G2 checkpoint function and is the most prognostic genelist (log-rank ! 2 = 8.5, p = 0.00351, Figure 4). We also observe that the misclassification rates based on PAM analysis are high for all the genelists. The misclassification rates are as follows: 36%,41%, 31% and 36% for the SAM, LIMMA, LPC and QTA methods, respectively (Table 3). 6. CONCLUSION

In this study, we compared four methods (SAM, LIMMA, LPC and QTA) for identifying DEGs in terms of their power to detect differential gene expression, the predictive ability of the genelists for a continuous outcome, and the prognostic properties of the genelists for DMFS. One simulated dataset and two publicly available datasets from melanoma studies were used

Figure 2: Number of overlapping genes from the SAM, the LIMMA, the LPC and the QTA genelists based on the melanoma cell lines dataset.

Table 2: Comparison of G2 Checkpoint Function Prediction by the SAM, LIMMA, LPC and QTA Genelists Generated at α = 0.1. The Number of Genes Associated with DMFS (Cox Genes) are also Included

Method # Genes in model r p R2 Adjusted R2 # Cox genes

SAM 10 0.652 <0.001 0.43 0.193 5 LIMMA 6 0.550 0.0006 0.3 0.150 1 LPC 3 0.421 0.0117 0.18 0.100 1 QTA 4 0.721 <0.001 0.52 0.456 1 Combine 16 0.710 <0.001 0.5 0.105 6

Figure 3: Kaplan-Meier curves for the low- and high-risk groups, generated by A. the SAM genelist, B. the LIMMA genelist, C. the LPC genelist, and D. the QTA genelist.

in this regard. Results show that the selection of the DEGs heavily depends on the choice of the gene selection method. This may be due to the assumptions made by different methods. The LIMMA method assumes that the null distribution of the test statistics is the same for all genes. The QTA approach depends heavily on the normality and linearity assumptions, and the SAM method, in case of two groups scenario, assumes equal variance. Therefore, to obtain the reliable results for detecting significant genes in microarray data analysis, we need to explore the characteristics of the data and then apply the most appropriate method under the given situation.

In addition to finding DEGs, it is also important to assess the biological and clinical importance of these genelists. One way of doing this is by identifying gene

signatures that are better predictors of a quantitative outcome or a patient's survival. This may help in tailoring therapeutic strategies to a single patient rather than the one-size-fits-all paradigm. Results from this study have shown that the combined genelist is more accurate in separating melanoma patients into high/low risk groups for developing distance metastasis. While the SAM approach was more powerful in terms of the number of significant genes detected using real dataset, the genelist generated by the QTA approach performed better in terms of prediction. Therefore, the QTA method would be preferred over the other approaches in predicting a quantitative outcome.

Omolo et al. [20] employed the QTA method (together with a Bayesian procedure) to identify 165 genes that were associated with the G2 checkpoint

function in melanoma lines. Some of these genes were found to be expressed differentially in wild-type (WT), NRAS-mutant and BRAF-mutant melanoma lines, through RNA expression analysis. This 165-list was also prognostic for distant metastasis-free survival in primary melanomas. Our SAM-list (n=33), LIMMA-list (n=22), LPC-list (n=7) and QTA-list (n=4) had ten (10), three (3), three (3) and one (1) genes in common with the 165 gene list, respectively. Kaufmann et al. [26] showed that some of the genes correlated with chromosomal instability (n=190), obtained using the QTA and a Bayesian method, were linked to amplification or deletion of the gene, e.g. DDR2. Our SAM-list and the LIMMA-list had two (2) genes each in common with the 190-list, which included DDR2. Thus, some of the DEGs by the proposed statistical methods in this manuscript have been biologically validated to be true positives (TP) in recent studies.

Future work should focus on the development of more methods for differential gene expression analysis, since none of the methods discussed in this work and other existing survey papers is recommended as the

“gold-standard". Our study was limited to the two microarray datasets from melanoma research, but the results would still hold when multiple datasets are considered.

ACKNOWLEDGEMENTS

We are indebted to Dr. William K. Kaufmann for the continuous and survival outcome data used in this study. This work was partially supported by a grant from the Simons Foundation (# 282714 to BO). LC was supported by a scholarship from the African Union. SUPPLEMENTAL MATERIALS The supplemental materials can be downloaded from the journal website along with the article. REFERENCES

[1] Sreekumar J, Jose KK. Statistical tests for identification of differentially expressed genes in cDNA microarray experiments. Indian J Biotechnol 2008; 7: 423-436. [2] Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially

Figure 4: Kaplan-Meier curves for the low- and high -risk groups generated by the combined genelist containing 52 unique genes.

Table 3: Misclassification Rates Based on the Prediction Analysis for Microarrays (PAM)

	SAM	LIMMA	LPC	QTA
Misclassification rate	36%	41%	31%	36%

expressed genes in microarray data. Bioinformatics 2002; 18: 1454-1461. https://doi.org/10.1093/bioinformatics/18.11.1454 [3] Schwender H, Krause A, Ickstadt K. Comparison of the empirical bayes and the significance analysis of microarrays. Technical Report//Universitt Dortmund, SFB 475, Reduction of complexity in multivariate data structures; 2003. [4] Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 2006; 7: 359. https://doi.org/10.1186/1471-2105-7-359 [5] Kim SY, Lee JW, Sohn IS. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. Stat Methods Med Res 2006; 15: 3-20. https://doi.org/10.1191/0962280206sm423oa [6] Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, Guedj M. Should we abandon the t

-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. PLoS One 2010; 5: e12336. [7] Bair E. Identification of significant features in DNA microarray data: Feature selection in DNA microarray data. Wiley Interdiscip Rev Comput Stat 2013; 5: 309-325. https://doi.org/10.1002/wics.1260 [8] Bandyopadhyay S, Mallik S, Mukhopadhyay A. A survey and comparative study of statistical tests for identifying differential expression from microarray data. IEEE/ACM Trans Comput Biol Bioinformatics 2014; 11: 95-115. https://doi.org/10.1109/TCBB.2013.147 [9] Kaufmann WK, Nevis KR, Qu P, Ibrahim JG, Zhou T, Zhou Y, et al. Defective cell cycle checkpoint functions in melanoma are associated with altered patterns of gene expression. J Invest Dermatol 2008; 128: 175-187. https://doi.org/10.1038/sj.jid.5700935 [10] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001; 98: 5116-5121. https://doi.org/10.1073/pnas.091062498 [11] Smyth GK. limma: Linear models for microarray data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, Eds. Bioinformatics and computational biology solutions using R and Bioconductor. Springer New York 2005; pp. 397420. [12] Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001; 96: 1151-1160. https://doi.org/10.1198/016214501753382129 [13] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3: 1-25. https://doi.org/10.2202/1544-6115.1027

[14] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNAsequencing and microarray studies. Nucleic Acids Res 2015; 43(7): e47. https://doi.org/10.1093/nar/gkv007 [15] Witten DM, Tibshirani R. Testing significance of features by lassoed principal components. Ann Appl Stat 2008; 2: 9861012. https://doi.org/10.1214/08-AOAS182 [16] Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-Array Tools. Cancer Inform 2007; 3: 11-17. [17] Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to highdimensional genomic data. J Stat Plan Inference 2004; 124: 379-398. https://doi.org/10.1016/S0378-3758(03)00211-8 [18] Golub GH, Van Loan CF. Matrix computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press; 1996. Available from: https://books.google.co.ke/books?id=mlOa7wPX6OYC. [19] Owzar K, Jung SH, Sen PK. A copula approach for detecting prognostic genes associated with survival outcome in microarray studies. Biometrics 2007; 63: 1089-1098. https://doi.org/10.1111/j.1541-0420.2007.00802.x [20] Omolo B, Carson C, Chu H, Zhou Y, Simpson DA, Hesse JE, et al. A prognostic signature of G2 checkpoint function in melanoma cell lines. Cell Cycle 2013; 12: 1071-1082. https://doi.org/10.4161/cc.24067 [21] Winnepenninckx V, Lazar V, Michiels S, Dessen P, Stas M, Alonso SR, et al. Gene expression profiling of primary cutaneous melanoma and clinical outcome. J Natl Cancer Inst 2006; 98: 472-482. https://doi.org/10.1093/jnci/djj103 [22] Tibshirani RJ. Regression shrinkage and selection via the LASSO. J Roy Statist Soc B 1996; 58(1): 267-288. [23] Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004; 2. https://doi.org/10.1371/journal.pbio.0020108 [24] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002; 99: 6567-6572. https://doi.org/10.1073/pnas.082099299 [25] Andrew H,

Florence G, Golum Kibria B. Methods for identifying differentially expressed genes: An empirical comparison. J Biom Biostat 2015; 6(5). [26] Kaufmann WK, Carson CC, Omolo B, Filgo AJ, Sambade MJ, Simpson DA, et al. Mechanisms of chromosomal instability in melanoma: Chromosomal Instability in Melanoma. Environ Mol Mutagen 2014; 55: 457-471. https://doi.org/10.1002/em.21859

Received on 01-02-2017 Accepted on 23-02-2017 Published on 28-02-2017

https://doi.org/10.6000/1929-6029.2017.06.01.1