universiteit
hasselt
UNIVERSITEIT VAN DE TOEKOMST

Maastricht University

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Bioinformatics

Masterproef

Classification and class prediction for different chemical structures using gene expression data

Promotor :
Prof. dr. Ziv SHKEDY
Prof. dr. Dan LIN

Victor Lih Jong

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

Maastricht University

universiteit
hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Bioinformatics

Masterproef

Classification and class prediction for different chemical structures using gene expression data

Promotor :
Prof. dr. Ziv SHKEDY
Prof. dr. Dan LIN

Victor Lih Jong

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics

Maastricht University

universiteit hasselt

UNIVERSITEIT VAN DE TOEKOMST

DEDICATION

I dedicate this work to;

my late grandmother Margaret NACHI OJONG

for her love to our entire family in general and her trust in me in particular. Grandma, I will never forget your food for taught "winners never quit and quitters never win". While I try as much as possible not to fail you on this sinful earth, I wish that your soul rest in perfect peace.

And my parents Chief NFOATAW David OJONG and Martha ENOW ASHU

for their immense efforts that brought me up to this level. I love you all.

ACKNOWLEDGMENT

This research work couldn't have seen the light of the day under another researcher if not of the financial support I got from my parents, brothers/sisters and most especially from Corina BOLLI, Sybil DINGA, Peter FLEUTSCH, Simon SCHMID and Ivo TEBA.

I am highly indebted to my supervisors Prof. Dr. Ziv SKHEDY and Prof. Dr. Dan LIN for their immense contributions from course work through their constant advice and guidance as regard to the best methods of handling this research work. I also wish to extend my profound gratitude to members of the staff at Center for Statistics, University of Hasselt, starting with Prof. Dr. Marc AERTS, chair Master of Epidemiology and Public Health Methodology, Prof. Dr. Geert MOLENBERGHS, chair Master of Biostatistics, Prof. Dr. Tomasz BURZYKOWSKI, chair Master of Statistical Bioinformatics and Mrs. Martine MACHIELS, Programs Manager; to mention just a few, for their great roles in making us part of the world's researchers.

My appreciations also go to my colleagues at Statistical Bioinformatics and other programs at center for statistics for their collective efforts from the start of our programs till now and also for the conducive atmosphere they created during our stay at center for statistics. I appreciate the contributions of Elvis NDAH, Toon VAN CAMPENHOUT, Patwary FAZLUL, Quinta NANGA, Anne-Marie SHUDZEKA, Adeline NDIFOR, Raoul OUANDJI, Katrien JEURIS, Geraldine AGBOR, Angelica MOLINA, Miranda AKATEH, Zemicael AKLILU, Yannick VANDENDIJCK, Francis BATOMEN, Geraldine NGWANA and Sisay TANIE for their contributions as my group members in one or two of the courses during this study period.

Last but not the least, I wish to thank all my friends most especially Clerinetus AGBOR, Roland AYUK, Collins AKOSA, Tom AKO, Becky ASHU, Brain TANYI, Jean Filbert NJONKWO, Kenneth NEMBU, Nature NJUNGWA, David AKWANGA and Calistus JONG for being there always to spice and relax the atmosphere when it was tensed. Special appreciation goes to Dr. NJI ABATIH E. as an all-round mentor. Thanks also go to everyone who has not been listed here but who supported me by any means during my studies. Our lord should bless you all abundantly.

SUMMARY

Background: For years microarray-based classification has been a major topic in statistics, bioinformatics and biomedical research but because of the large number of variables as compared to the sample size, traditional statistical methods have been unsatisfactory. Thus, special methods for microarray data together with data mining technologies have been developed to address these unsatisfactory issues. The aim of this project was to apply some of these technologies to build classification function(s) that can be used to classify chemical compounds to their identified clusters and also to predict the cluster of a new chemical compound based on gene expression data. The data contained sixty chemical compounds grouped into three clusters GC14, GC22 and GC29 with the clusters having 32, 13 and 15 chemical compounds respectively.

Results: Analysis of differentially expressed genes showed that the three clusters were significantly different at 5% significance level as was portrayed by over 500 rejected hypotheses using parametric, nonparametric and resampling-based procedures. Though the three clusters were found to be different, contrasts analyses showed that cluster GC22 and GC29 were closely similar as there were no rejected hypotheses from the parametric and nonparametric procedures and just barely a few rejected hypotheses from the resampling-based procedure, at 5% significance level. Also, a classification analysis yielded a misclassification rate of approximately 32% when the three clusters were considered separately and a misclassification rate of approximately 7%, a sensitivity of 89% and specificity of 96% when the two closely similar clusters were merged into one cluster.

Conclusion: Clusters GC22 and GC29 were found to be indifferent and as such should be considered as a single cluster. And for this setting, the combination of {10221_at, 100287237_at, 100141515_at, 10095_at, 100131187_at, 100128071_at, 100286909_at, 100287098_at, 10217_at, 100129637_at} as gene signature with linear discriminant analysis as a classification function will predict the cluster of an existing or new chemical compound with very low misclassification chances.

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background

The identification of functional causes of a disease is usually a primary concern of biomedical research. To understand the contributing mechanism of a disease, the desire has been to identify genes that are associated with such a disease. Gene expression is measured in several ways including mRNA and protein expression. Thus, the purpose of such research is to answer the questions: “which genes are associated with a tumour?”, “which clusters of genes are involved in a particular tumour?” and/or “to which tumour class does a particular sample belong?”.

As one of the rapidly growing technologies in the field of genomics, the microarray technology is widely used for the analysis of gene expression data. Due to the large amount of data generated by the microarray experiments in terms of the parameters as compared to the sample sizes (p>>n), traditional statistical methods have become handicap to explore these large data sets. As such, special statistical methodologies for microarray data together with some data mining technologies have been brought together to automatically tackle these large data sets.

In this study some of these methods have been implemented to analyse data generated from one of the microarray (Affymetrix) technologies with the intention to identify a set of genes that can clearly distinguish between clusters of chemical compounds used at early stage of drug discovery. One of the benefits of this study might be to identify which (cluster of) chemical compound(s) may respond well to a particular disease based on the associated genes to this disease/chemical compound through their expression levels.

The following sub-section describes briefly the data and the objective of the research, section two presents the methodologies implemented in the analysis. This includes methods used to identify differentially expressed genes, construct classification functions and build class predictors. The results of the analysis are presented in section three, section four presents a brief discussion while section five described briefly the software, packages and functions used for the entire analysis. The references are presented in section six while section seven is the appendix and contains some figures from the results of the analysis.

1.2. Data and Objectives

The data set is composed of three clusters of chemical compounds used in the early stage of drug discovery. These chemical compounds were grouped into clusters depending on their chemical structures, chemical formulae and other chemical/physical characteristics. Each chemical compound was then applied to a cultured specimen and the expression levels of the genes measured. For this project, the clusters used were namely GC14, GC22 and GC29. Cluster GC14 is composed of 32 chemical compounds while GC22 is composed of 13 chemical compounds and cluster GC29 is composed of 15 chemical compounds. After normalization and gene filtering, the total number of genes left for which expression levels were used, was 7722 genes.

In order to address the research interest, the following scientific questions were posed: Are the three clusters of chemical compounds different? If yes, where does the difference lies? Is it possible to select a reasonable number of genes for which a classification function can be built to classify existing chemical compounds to their respective classes (clusters) and/or predict the class (cluster) of a new chemical compound with acceptable error rates? To answer these scientific questions, the objectives of the study were reformulated to assess firstly if there are any differentially expressed genes between the three clusters and if there are any, the contrasts are investigated to clearly identify where the differential expression arises. And secondly to build a classifier that can correctly classify an existing chemical compound to its cluster and/or predict the cluster of a new chemical compound with a tolerated amount of error, using the expression levels of the genes.

2. METHODOLOGY

2.1. Differentially expressed genes

2.1.1. Three clusters analysis

Let $\mu_{iA}, \mu_{iB}, and\ \mu_{iC}, i$=1, 2, ..., 7722 be the mean expressions for gene *i* in the three clusters GC14, GC22 and GC29 respectively. To assess if the three clusters are the same, we test for each and every gene the following hypothesis:

$$H_{0i}\ :\ \mu_{iA} = \mu_{iB} = \mu_{iC} \qquad versus\ \ H_{1i}\ :\ not\ all\ \mu_i are\ equal$$

This comparison was done using the following statistical tests:

- ✓ **F test**

Let Y_{ijk} be the expression level of gene *i* in cluster *j* for subject *k*, $Y_{ij.}$be the expression mean of gene *i* in cluster *j* and $Y_{i..}$ be the overall expression mean of gene *i*. To test for the above hypothesis we fit the model:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad (1)$$

where *j= A, B, C* represents the different clusters and *k= 1, 2, ...,* n_j represents the number of chemical compounds in cluster j with $E(Y_{ijk}) = \mu_{ij} = Y_{ij.}$. The assumptions of model (1) are that the epsilons are independently normally distributed errors with mean zero and constant variance that is $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

The test statistic for model (1) above is

$$F_i = \frac{\frac{\sum_j n_j (Y_{ij.} - Y_{i..})^2}{(r-1)}}{\frac{\sum_j \sum_k (Y_{ijk} - Y_{ij.})^2}{(n_T - r)}} \sim F(r-1, n_T - r)$$

Where r is the total number of clusters and n_T is the total number of chemical compounds in all the clusters. For a significance level α, the statistic follows an F distribution with *r-1* and $n_T - r$ degrees of freedom. Thus, we reject H_0 if $F_i > F(1-\alpha, r-1, n_T - r)$ and fail to reject H_0

otherwise (Kutner et al., 2005). The numerator of the statistics is the between group (cluster) variability while the denominator is the within group (cluster) variability.

- ✓ **Moderated F test**

The moderated F test is similar to the F test but for the fact that the test statistic does not make any constant assumption about variance between genes. More specifically, the moderated F test model is as follows:

$$Y = X\beta + \varepsilon \qquad (2)$$

Where Y is the expression matrix, X is the design matrix and β is the contrast matrix with each row of the design matrix corresponding to an array in the experiment and each column corresponds to a coefficient which is used to describe the RNA sources in the experiment (Smyth, 2004). Unlike the F test which assumes that the variability is constant between genes, the moderated F test assumes prior distributions of the model coefficients and the variances of the genes and then computes the posteriors estimates of these parameters through the empirical Bayes approach. The final gene specific within group variance is then computed as a weighted average of the posterior estimate of the overall variability s_0^2 and per gene variability s_g^2. This renders a similar test statistic as the F statistic but with augmented degrees of freedom (Smyth, 2004).

- ✓ **Kruskal-Wallis test**

For each gene *i*, the expression levels are replaced with ranks and let $R_{ij.}$ be the average rank of gene *i* for cluster j, *j*=1,...,s and $R_{i..}$ the overall average rank for gene *i* then the clusters differ widely among each other if and only if there are big differences among the values of the $R_{ij.}s$ (Lehmann, E. L., 2006). A convenient measure of the overall closeness of $R_{ij.}$ to $R_{i..}$ is a weighted sum of square difference defined below and known as the *Kruskal-Wallis* statistic

$$K = \frac{12}{N(N+1)}\sum_{j=1}^{s} n_j (R_{ij.} - R_{i..})^2$$

where n_j are the number of chemical compounds in cluster j, *N* total number of chemical compounds and s is the total number of clusters. K is zero when $R_{ij.}$ are equal and is large when there are substantial differences among the $R_{ij.}$ (Lehmann, E. L., 2006). We then reject the null hypothesis at a given significance level α if

$$P_{H_o}[K \geq k] \leq \alpha$$

where *k* is the observed value for *K*. Lehmann, E. L. (2006) states that for three or more clusters and cluster sample sizes greater than or equal to five, the distribution of the statistics is approximated by the Chi-square distribution and $P_{H_o}[K \geq k] \approx \Psi_{s-1}(k)$.

✓ **Significance Analysis of Microarrays (SAM)**

SAM is a resampling-based procedure in which no distributional assumption is made. It uses permutations to approximate the null distribution of the test statistic. The SAM test statistics is as follows:

$$F_i^* = \frac{\frac{\sum_j n_j (Y_{ij.} - Y_{i..})^2}{(r-1)}}{\frac{\sum_j \sum_k (Y_{ijk} - Y_{ij.})^2}{(n_T - r)} + s_0}$$

The constant s_0 is called the fudge factor and it is estimated as the percentile of the gene-wise standard errors that minimizes the coefficient of variation of the SAM test statistics. This modification is used to overcome bias for genes with expression difference (between clusters variability) close to zero, which have large values of the test statistics due to small within clusters variances. Supposed *B* permutations are made, the SAM matrix of statistics will be

$$F^{SAM} = \begin{pmatrix} F_{11}^* & F_{12}^* & \dots & F_{1B}^* \\ F_{21}^* & F_{22}^* & \dots & F_{2B}^* \\ \vdots & \vdots & \vdots & \vdots \\ F_{m1}^* & F_{m2}^* & \dots & F_{mB}^* \end{pmatrix}$$

where 1, 2, ... *m* rows are the genes and 1, 2, ..., B columns are the permutations. Once this matrix is obtained, the columns are sorted and the row means are computed to yield the expected statistics

$$\overline{F^{SAM}} = \begin{pmatrix} \bar{F}_1 \\ \bar{F}_2 \\ \vdots \\ \bar{F}_m \end{pmatrix}$$

To call a gene significant, the difference between the observed and expected values of the test statistic needs to be larger than a certain cut-off value λ. For a grid of λ values, the corresponding number of significant genes can be listed; at the same time, the number of false positives arising from any permutation matrix F^{SAM} is estimated. Under the null hypotheses, we expect that no differentially expressed genes are present for each permutation. Consequently the median or 90 percentile number of false positives corresponding to λ can be obtained from permutation matrix. In this way, the FDR can be calculated for each value of λ and an acceptable value of λ can be chosen to control the FDR at the desired level. (Lin et al., 2008).

The SAM Procedure also controls for false discovery rate (FDR) once the permutation matrix is obtained. Apart from automatically controlling for multiple testing, SAM has the strength that the null distribution is generated for all the genes at once by permuting the group labels, so that the correlation between test statistics of all the genes is preserved (Lin et al., 2008).

2.1.2. Contrasts analysis

Once a reasonable number of the null hypotheses in the three cluster analysis are rejected signifying a difference between the groups, analysis of the contrasts GC14-GC22, GC14-GC29 and GC22-GC29 is carried out to identify where the difference lies. In each of these contrasts and for each and every gene *i*, the following respective hypotheses were formulated:

$$H_{0i} : \mu_{iA} - \mu_{\mathrm{iB}} = 0 \quad versus\ H_{1i} : \mu_{iA} - \mu_{iB} \neq 0$$

$$H_{0i} : \mu_{iA} - \mu_{\mathrm{iC}} = 0 \quad versus\ H_{1i} : \mu_{iA} - \mu_{iC} \neq 0$$

$$H_{0i} : \mu_{iB} - \mu_{\mathrm{iC}} = 0 \quad versus\ H_{1i} : \mu_{iB} - \mu_{iC} \neq 0$$

These comparisons were done using the following statistical tests:

- ✓ **t test**

Let Y_{ijk} be the expression level of gene *i (i =1, ..., m)* in cluster *j* for subject *k,* and μ_{ij} be the expression mean of gene *i* in cluster *j*. The test statistics for the hypotheses can be written as

$$t_i = \frac{\mu_{ij} - \mu_{il}}{s_i\sqrt{\frac{1}{n_j} + \frac{1}{n_l}}} \quad i = 1, \dots, m, \quad j, l = A, B, C \; and \; j \neq l$$

where $s_i^2 = \frac{1}{n_j+n_l-2}\{\sum_{k=0}^{n_j}\left(Y_{ijk} - \mu_{ij}\right)^2 + \sum_{k=0}^{n_l}\left(Y_{ijk} - \mu_{il}\right)^2\}$ is the pooled variance for gene *i*, n_j is number of arrays (chemical compounds) in cluster *j* and n_l is the number of arrays (chemical compounds) in cluster *l.* This statistic has the assumptions that the samples are randomly and independently assigned to the clusters, the variance for gene *i* is the same across the clusters and that the population from which the samples are selected both have approximately normal relative frequency distribution. Thus this statistics follows a t distribution with $n_j + n_l - 2$ degrees of freedom. (Mendenhall & Sincich, 2007).

- ✓ **Moderated t test**

Like with the moderated F test, a linear model is fitted for each gene and the null hypothesis of zero coefficients is investigated using the t test except of the fact that the test statistic does not make any constant assumption about variance of the error terms. Unlike the t test, the moderated t test assumes prior distributions of the model coefficients and the variances of the error terms and then computes the posteriors estimates of these parameters through the empirical Bayes approach. The final gene specific within group variance is then computed as a weighted average of the posterior estimate of the overall (prior) variability s_0^2 and observed per gene variability s_g^2. The posterior values shrink the observed variances towards the prior values with the degree of shrinkage depending on the relative sizes of the observed and prior degrees of freedom. Under the null hypothesis of zero coefficients, it yields a statistic called the moderated t statistic that has a t distribution with degree of freedom as the sum of the prior and the observed degrees of freedom (Smyth, 2004).

✓ **Wilcoxon Rank Sum test**

For each of the hypotheses above, let j and l be the two clusters to be tested for equal means with *m* and n the number of chemical compounds respectively in each cluster. For each gene i, the expression values are assigned ranks and let W_{ij} and W_{il} be the sum of ranks of gene i in clusters j and l respectively then for gene i, we reject the null hypothesis at a given significance level α if

$$P_{H_o}\left[\left|W_{il}-\tfrac{1}{2}n(N+1)\right|\geq\left|w_i-\tfrac{1}{2}n(N+1)\right|\right]\leq\alpha$$

Where w_i is the observed value of W_{il}, N is the total number of chemical compounds in the two clusters. Lehmann, E. L. (2006) states for large values of *m* and *n* the distribution of the statistics is approximated by that of the normal distribution and the significance of the tow-sided test hypothesis is then approximated by

$$2\left\{1-\Phi\left[\frac{\left|w_i-\frac{1}{2}n(N+1)\right|-\frac{1}{2}}{\sqrt{\frac{mn(N+1)}{12}}}\right]\right\}$$

✓ **Significance Analysis of Microarrays (SAM) test**

In a similar manner as for the t test, let Y_{ijk} be the expression level of gene *i (i =1, ..., m)* in cluster *j* for subject *k*, and μ_{ij} be the expression mean of gene *i* in cluster *j*. The test statistics for the hypotheses can be written as

$$t_i^*=\frac{\mu_{ij}-\mu_{il}}{s_i\sqrt{\frac{1}{n_j}+\frac{1}{n_l}}+s_0}\qquad i=1,\dots,m,\quad j,l=A,B,C\ and\ j\neq l$$

where $s_i^2=\frac{1}{n_j+n_l-2}\{\sum_{k=0}^{n_j}\left(Y_{ijk}-\mu_{ij}\right)^2+\sum_{k=0}^{n_l}\left(Y_{ijk}-\mu_{il}\right)^2\}$ is the pooled variance for gene *i*, n_j is number of arrays (chemical compounds) in cluster *j* and n_l is the number of arrays (chemical compounds) in cluster *l*. The constant s_0 is called the fudge factor and it is estimated as the percentile of the gene-wise standard errors that minimizes the coefficient of variation of the SAM test statistics. This modification is used to overcome bias for genes with expression difference

$\mu_{ij} - \mu_{il}$ close to zero having large values of the test statistics due to small sample variances. Supposed B permutations are made, the SAM statistics will be

$$T^{SAM} = \begin{pmatrix} t_{11}^{*} & t_{12}^{*} & \dots & t_{1B}^{*} \\ t_{21}^{*} & T_{22}^{*} & \dots & T_{2B}^{*} \\ \vdots & \vdots & \vdots & \vdots \\ . & . & . & . \\ t_{m1}^{*} & t_{m2}^{*} & \dots & t_{mB}^{*} \end{pmatrix}$$

where 1, 2, ... m rows are the genes and 1, 2, ..., B columns are the permutations. Once this matrix is obtained, the analysis follows suit as in the multiclass case described above.

2.1.3. Controlling the FDR for multiple testing

In all cases except for SAM procedures, I corrected for multiple testing by controlling the false discovery rate (FDR) defined as expected proportion of false rejection among the rejected hypotheses using the Benjamini and Hochberg (BH) procedure described by Lin et al. (2010). The BH procedure is a linear step-up procedure in which if the desired FDR level is α then the ordered p-value $P_{(i)}$ is compared to the critical value $\alpha \times \frac{i}{m}$. Let $k = max\{i: P_{(i)} \leq \alpha \times \frac{i}{m}\}$, then reject $H_{(1),} \dots, H_{(k)}$ if such a k exists (Lin et al., 2010) .

2.2. Classification

Let $X \subset \mathbb{R}^P$ be the predictor space, $Y = \{0,1, \dots K-1\}$ be the vector of finite (K) set of class labels and $P(X,Y)$ be the joint probability distribution on $X \times Y$. Let also $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be a sample of n predictor-class pairs. Then the classification task is to construct a decision function

$\hat{f}: X \rightarrow Y$ (Where ^ indicates that the function is estimated from the sample.)
$x \rightarrow \hat{f}(x)$

such that the generalization error

$$R[f] = E_P[L(f(x), y)] = \int_{X \times Y} L(f(x), y) dP(x, y)$$

is minimized. Where $L(.,.)$ is a suitable loss function with $L(u,v) = 1 \; if \; u \neq v$, $L(u,v) = 0 \; otherwise$ (Slawski et al., 2008).

2.2.1. Performance measures

✓ **Misclassification error**

Since we are only equipped with a finite sample and the underlying distribution is unknown, the empirical counterpart to the generalization error is estimated as

$$R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$$

Though this empirical counterpart to the generalization error can be used to evaluate classifiers, it usually overfits the sample S. Thus a general practice is to split the sample into a learning set L and a test set T and $\hat{f}(.)$ is constructed from L only and evaluated using T (Slawski et al., 2008).

Because the sample sizes are usually very small, a good practice is to generate several learning and test sets from the available sample, construct a classifier with each learning set and using the corresponding test set estimate the empirical generalization error. The final empirical generalization error will be the average across the test set. Suppose B learning sets $\mathrm{L_b}$ $(\mathrm{b} = 1, \dots \mathrm{B})$ are generated from S and the corresponding test set $\mathrm{T_b} = \mathrm{S} \backslash \mathrm{L_b}$ with $\hat{\mathrm{f}}_{\mathrm{b}}(.)$ obtained from $\mathrm{L_b}$ $(\mathrm{b} = 1, \dots \mathrm{B})$ then an estimate of the error rate is

$$\hat{\mathrm{E}} = \frac{1}{\mathrm{B}}\sum_{\mathrm{b}=1}^{\mathrm{B}} \frac{1}{|\mathrm{T_b}|}\sum_{\mathrm{i}\epsilon \mathrm{T_b}} \mathrm{L}(\mathrm{y_i}\ \hat{\mathrm{f}}_{\mathrm{b}}(\mathrm{x_i}))$$

where $|\cdot|$ is the cardinality of the considered set (Slawski et al., 2008).

✓ **Sensitivity and specificity**

For binary classifications, the following measures were also used alongside the misclassification error. Let the two classes be + and - for notation purpose and consider the following table which contains the true class (cluster) and predicted class (cluster) of an observation (chemical compound)

Predicted class of observation (T)

		+	−
True class of observation (S)	+	*a*	*b*
	−	*c*	*d*

We define sensitivity as the probability that the predicted class is + given that the true class is + that is $P(T = +|S = +) = \frac{a}{a+b}$ and specificity as the probability that the predicted class is $-$ given that the true class is $-$ that is $(T = -|S = -) = \frac{d}{c+d}$. Higher values of sensitivity and specificity entail a good classifier (Agresti et al, 1997).

2.2.2. Features selection for three clusters analysis

For 2000 bootstrap, the data was split into $\frac{2}{3}$ for the training set and $\frac{1}{3}$ for the test set taking into account the number of samples per cluster using Monte Carlo Cross Validation (MCCV) described by Slawski et al. (2008). For each bootstrap, classifiers were built from a combination of one of F test, moderated F test and Kruskal-Wallis test as gene selection method, five of the classification functions described below and top k genes where k = 5, 10 and 20. These classifiers were evaluated with each bootstrap test set and the misclassification errors recorded. The means and the standard errors of the misclassification errors over the 2000 bootstrap for the classifiers built from each combination were computed and the combination with the lowest misclassification error and/or smallest standard error was chosen for further analysis. Let the combination with lowest misclassification error be denoted $method3^*,\ classifier3^*\ and\ topk3^*$.

2.2.3. Features selection for two clusters analysis

In a similar manner, for 2000 bootstrap, the data was split into $\frac{2}{3}$ for the training set and $\frac{1}{3}$ for the test set taking into account the number of samples per cluster using Monte Carlo Cross Validation (MCCV) described by Slawski et al. (2008). For each bootstrap, classifiers were built from a combination of one of t test, moderated t test and Wilcoxon rank sum test as gene selection method, five of the classification functions described below and top k genes where k = 5, 10 and 20. These classifiers were evaluated with each bootstrap test set and the misclassification errors recorded. The means and the standard errors of the misclassification errors over the 2000 bootstrap for the classifiers built from each combination were computed and the combination with the lowest misclassification error and/or smallest standard error was chosen for further analysis. Let the combination with the lowest misclassification error be $method2^{*}, \ classifier2^{*} \ and \ topk2^{*}$.

2.2.4. Gene signatures and final classifiers

In each clusters (three or two) setting and for 2000 bootstraps the data was split into $\frac{2}{3}$ for the training set and $\frac{1}{3}$ for the test set and using $method3^{*} \ and \ topk3^{*}$; $method2^{*} and \ topk2^{*}$ for three clusters setting and two clusters setting respectively, $topk3^{*} \ and \ topk2^{*}$ genes that were selected most of the times were extracted as gene signatures for three clusters setting and two clusters setting respectively. Finally, for 2000 iterations the data was split in a similar manner as above into $\frac{2}{3}$ for the training set and $\frac{1}{3}$ for the test set and for each gene signature say $sig3^{*} \ or \ sig2^{*}$ for three clusters and two clusters settings respectively, classifiers were built using the classification functions $classifier3^{*} \ \ and \ \ classifier2^{*}$ respectively for the three and two clusters settings. In each setting, misclassification rate and/or sensitivity–specificity analysis was carried out.

2.2.5. Classification functions (classifiers)

A. Linear discriminant analysis (LDA) and/or Diagonal linear discriminant analysis (DLDA)

Let *G* be a vector of class labels and *X* a matrix of covariates. Suppose $f_k(x)$ is the class-conditional density of *X* in class *G* = *k*, and let π_k be the prior probability of class *k*, with $\sum_{k=1}^{K} \pi_k = 1$, the goal of classification by discriminant analysis is to estimate the posterior probability

$$P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Thus, having the posterior probabilities uses techniques that are based on models for the class densities (Hastie et al., 2009). Discriminant analyses are Bayes optimal classifiers which assume that the conditional distributions of predictors given the classes are Gaussian (Slawski et al., 2008). They differ only by the assumptions made for their covariance matrices. LDA assumes that the within-class covariance matrices are equal for all the classes while DLDA assumes that the within-class covariance matrices are diagonal and equal for all classes.

B. K-nearest neighbor (KNN)

For a test observation x, the k-nearest neighbor classifier classifies this observation x based on a measure of distance between x and other training observations. It finds the k observations in the learning set closest to x and then predicts the class of x by majority votes. The value k is usually specified by the user but it should be noted that if k is too small, then the nearest-neighbor classifier may be susceptible to over-fitting because the noise in the training data. On the other hand, if k is too large, the nearest-neighbor classifier may misclassify the test instance because its list of nearest neighbors may include data that are located far away from its neighborhood (Tan et al., 2005). The optimal value of k can be chosen by cross validating and returning the one with the smallest misclassification error.

C. Random forest (RF)

Random forest is a classification method designed for decision tree classifiers. It combines the prediction made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors (Tan et al., 2005). Randomization helps to reduce the correlations among decision trees so that the generalization error of the classifier can be improved. A random vector can be incorporated in the tree-growing process in many ways some of which include; randomly select a subset (e.g. square root of total number) of the features and then grow a tree to its entirety or randomly select one of the best split at each node of the decision tree. Once multiple trees have been built, they are then combined by voting; that is each tree cast a vote at its terminal node.

D. Tree-based boosting (TBB)

Boosting is a classification method that combines the output of several “weak” classifiers to produce a powerful “committee” (Hastie et al., 2009). It is an iterative procedure used to adaptively change the distribution of the training examples so that the base classifiers will focus on examples that are hard to classify. Boosting assigns a weight to each training example and may adaptively change the weight at the end of each boosting round (Tan et al., 2005). These weights are then used either as a sampling distribution or can be used by the base classifier to learn a model that is biased toward higher-weight examples. The idea is to give all observations same weights at the start, perform a bootstrap sample and build a classifier in this case a classification tree (hence tree-based boosting) then test the classifier with all the objects. The weights of misclassified objects are increased in the next bootstrap sample thereby given them higher chances to be sampled.

E. Support vector machines (SVM)

Support Vector Machines classification is a binary classification method whereby it fits an optimal hyperplane between the two classes by maximizing the margin between the classes' closest points. The points lying on the boundaries are called support vectors, and the middle of the margin is the optimal separating hyperplane. Data points on the “wrong” side of the discriminant margin are weighted down to reduce their influence. For nonlinear cases, SVM uses a nonlinear mapping (via kernels) to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane that is, a “decision boundary” separating the tuples of one class from another. The SVM finds this hyperplane using support vectors (“essential” training tuples) and margins defined by the support vectors (Han, J. & Kamber, M., 2006). For multiclass classification, SVMs uses one-against-one technique by fitting all binary subclassifiers and finding the correct class by a voting mechanism (Meyer, 2011).

F. Neural Networks (NNET)

A neural network is a two-stage regression or classification model. The central idea is to extract linear combinations of the input variables as derived features, and then model the target as a nonlinear function of these features. This transformation can be done more than once leading to multi-hidden-layers neural networks but in this project, a one-hidden-layer neural network called the feed-forward neural networks was implemented. The idea is that; starting with covariates X, one forms projections $Z_r = \sigma(a_r^T X); r = 1, \dots, R$ which forms the units of the one-hidden-layer. These units are subsequently used as inputs for the prediction model. Here the activation function $\sigma(v)$ is usually chosen to be sigmoid that is $\sigma(v) = \frac{1}{1+e^{-v}}$ (Hastie et al., 2009).

3. RESULTS

This section provides the results obtained from the analysis of the dataset. The first subsection presents the differentially expressed genes found for the three clusters and their contrasts while the other subsection presents the classification results for the three clusters and two clusters formed by merging clusters GC22 and GC29 to one cluster.

3.1. Differentially expressed genes

Figures A1(a) and A1(b) in the appendix illustrate the need of SAM analysis. From these figures, one notices that there are test statistics with large values caused by small within clusters variability for the F test or sample standard errors for the t test but with very small between clusters variability or fold change respectively.

3.1.1. Three clusters analysis

Table 1 below shows the number of differentially expressed genes by each statistical test at different significance levels before and after adjusting for multiple testing. From this table one clearly sees that irrespective of the statistical test employed, there is high evidence that at least two of three clusters are different as portrayed by the number of rejected hypotheses.

Table 1: Number of rejected hypotheses by each test statistic at different significance levels

Statistical method	Alpha	Raw p-value	BH adjusted p value
F test	1%	1297	568
	5%	2320	1169
	10%	3054	1669
Moderated F test (limma)	1%	1268	555
	5%	2298	1162
	10%	3021	1563
Kruskal-Wallis test	1%	1751	873
	5%	2883	1843
	10%	3548	2510
SAM (S_0=0.0015)	1%	1340	546
	5%	2408	1039
	10%	3130	1478

For each test statistic, the top ten most significant genes at a significance level of 5% were retrieved and are shown in Table 2 below. From this table, one clearly sees that none of the genes appeared in more than one of the methods implying the methods have different lists of top 10 genes.

Table 2: Top 10 significant genes for each test statistic at 5% significance level.

	F test	Moderated F test	Kruskal-Wallis	SAM
1	152006_at	6156_at	7549_at	9314_at
2	135932_at	7388_at	3068_at	27154_at
3	64710_at	6194_at	51075_at	57602_at
4	5867_at	6165_at	143098_at	150094_at
5	8131_at	5692_at	79414_at	1962_at
6	27345_at	10767_at	26100_at	10221_at
7	23287_at	23516_at	339457_at	1846_at
8	220001_at	8667_at	80319_at	92335_at
9	26958_at	26476_at	94134_at	27042_at
10	285381_at	4883_at	388753_at	23406_at

3.1.2. Contrasts analysis

In order to investigate where the difference lies between the three clusters, contrasts were made and Table 3 below gives the number of rejected hypotheses for each and every contrast by each test statistic at 5% significance level before and after adjusting for multiple testing. Based on the results displayed on this table, one clearly sees that cluster GC14 is different from clusters GC22 and GC29 but cluster GC22 and GC29 seems to be the same because there is approximately zero differentially expressed gene between these two supposed clusters.

Table 3: Number of rejected hypotheses for the different contrasts at 5% significance level.

Contrast	Statistical method	Raw p-value	BH adjusted p-value
GC14 - GC22	t test	2499	1529
	Moderated t test (limma)	2256	1286
	Wilcoxon test	2933	1911
	SAM (S_0=0.0011)	2676	1390
GC14- GC29	t test	1576	384
	Moderated t test (limma)	1557	406
	Wilcoxon test	1840	718
	SAM (S_0=0.0011)	1613	347

GC22 – GC29	t test	848	0
	Moderated t test (limma)	1103	0
	Wilcoxon test	859	0
	SAM (S_0=0.0014)	864	3

As for the three clusters setting, the top 10 most significant genes for each of the statistical tests are presented in Tables 4 and 5 for the contrasts GC14 - GC22 and GC14 - GC29 respectively. From these tables one notices that for these contrasts, some genes appear amongst the top 10 of different statistical methods and that for some statistical methods, similar genes are differentially expressed between GC14 and GC22 or GC29, indicating that GC22 and GC29 are similar.

Table 4: Top 10 significant genes for GC14- GC22 by each test statistic at 5% significance level.

	t test	**Moderated t test**	**Wilcoxon test**	**SAM**
1	79026_at	150094_at	6461_at	9582_at
2	23244_at	57602_at	8723_at	253980_at
3	2831_at	9314_at	53838_at	65983_at
4	261726_at	1846_at	79738_at	56270_at
5	1106_at	27154_at	9966_at	90007_at
6	22828_at	1962_at	729970_at	8796_at
7	23300_at	10221_at	1106_at	10915_at
8	81853_at	92335_at	57106_at	161742_at
9	55858_at	79170_at	23483_at	25980_at
10	729970_at	84274_at	8021_at	163702_at

Table 5: Top 10 significant genes for GC14- GC29 by each test statistic at 5% significance level.

	t test	**Moderated t test**	**Wilcoxon test**	**SAM**
1	64782_at	9314_at	79894_at	3725_at
2	8611_at	1962_at	55341_at	339983_at
3	160897_at	27154_at	54841_at	9070_at
4	85455_at	113828_at	83607_at	63950_at
5	8773_at	57820_at	2152_at	83607_at
6	6137_at	23406_at	55147_at	7555_at
7	390502_at	27042_at	59_at	79074_at
8	66005_at	8140_at	152518_at	6574_at
9	9080_at	57602_at	81930_at	10486_at
10	22879_at	150094_at	5058_at	83734_at

3.2. Classification

3.2.1. Three clusters analysis

For 2000 bootstraps, features selection for the three clusters dataset was performed as described in the methodology and table 6 below gives the mean misclassification errors with standard errors in brackets, for all the various combinations of selection methods, top k genes and the classification functions. From this table, one notices that the combination of top 10 genes moderated F test as selection method and the classification functions LDA and NNET have the smallest misclassification error of 0.2701 with the same standard error of 0.0585.

Table 6: Three clusters mean misclassification error rates and their standard errors in brackets.

Top	Selection Method	SVM	DLDA	LDA	KNN	NNET
5	F test	0.3032(0.0628)	0.3601(0.0575)	0.2764(0.0612)	0.3601(0.0575)	0.2764(0.0612)
	Krusskal-Wallis test	0.3112(0.0631)	0.3921(0.0487)	0.2850(0.0575)	0.3921(0.0487)	0.2850(0.0575)
	Moderated F test	0.2991(0.0622)	0.3609(0.0595)	0.2726(0.0622)	0.3609(0.0595)	0.2726(0.0622)
10	F test	0.3007(0.0578)	0.3712(0.0541)	0.2712(0.0563)	0.3712(0.0541)	0.2712(0.0563)
	Krusskal-Wallis test	0.3083(0.0608)	0.3803(0.0485)	0.2764(0.0592)	0.3803(0.0485)	0.2764(0.0592)
	Moderated F test	0.3002(0.0591)	0.3735(0.0526)	**0.2701(0.0585)**	0.3735(0.0526)	**0.2701(0.0585)**
20	F test	0.3015(0.0569)	0.3539(0.0439)	0.3175(0.0638)	0.3539(0.0439)	0.3175(0.0638)
	Krusskal-Wallis test	0.2979(0.0596)	0.3528(0.0459)	0.3036(0.0649)	0.3528(0.0459)	0.3036(0.0649)
	Moderated F test	0.3021(0.0596)	0.3525(0.0440)	0.3186(0.0644)	0.3525(0.0440)	0.3186(0.0644)

Figure 1 below also represents the mean misclassification errors as a function of the top k genes and the selection methods. Based on this plot, the mean misclassification errors and their standard errors, the combination of top 10, moderated F test and LDA was chosen for further analysis. LDA could be replaced with NNET because they have the same misclassification error rates.

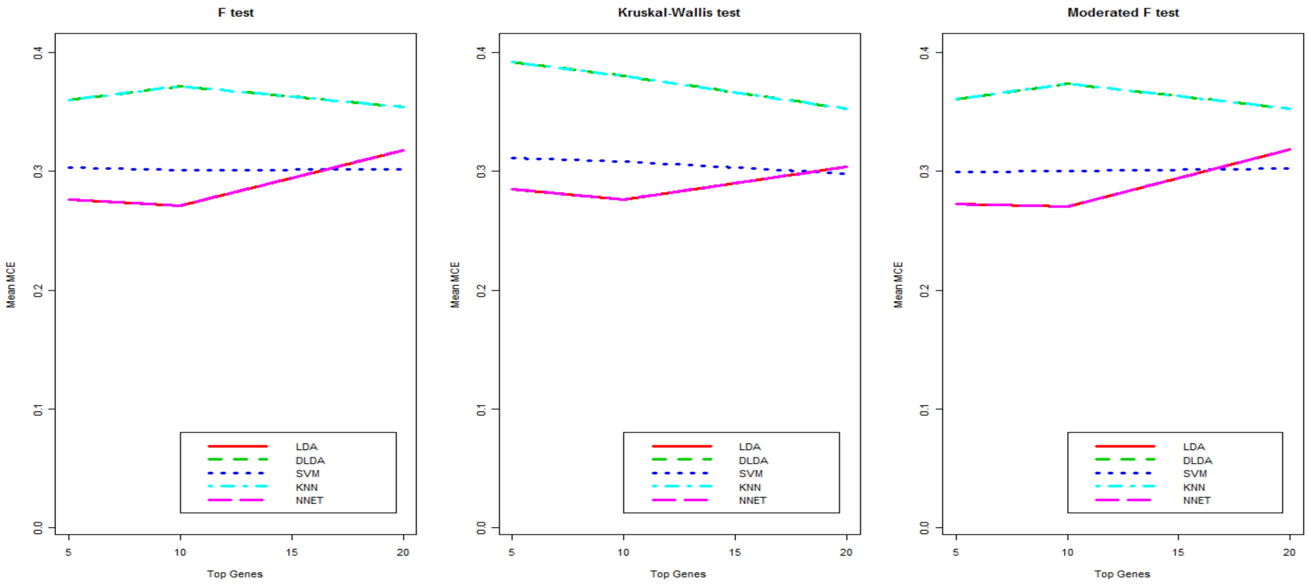


Figure 1: Mean misclassification error as a function of top k and selection method for 3 clusters.

For the chosen combination, 2000 bootstraps were performed for gene selection and table 7 below presents a list of the top 10 genes and their frequencies that were selected most of the times.

Table 7: Gene signature for the three clusters

Gene	10300_at	100288152_at	10152_at	100131512_at	10022_at	100127891_at	10018_at	10294_at	100288092_at	X10200_at
Freq.	1960	1799	1789	1707	1649	927	802	697	605	439

For this gene signature, the expression levels for each gene were compared by visualising a gene by gene box plot across the three clusters as shown in figure 2 below. From this figure, one clearly sees that the expression levels of the genes differ for GC14 and others but closely similar for GC22 and GC29.

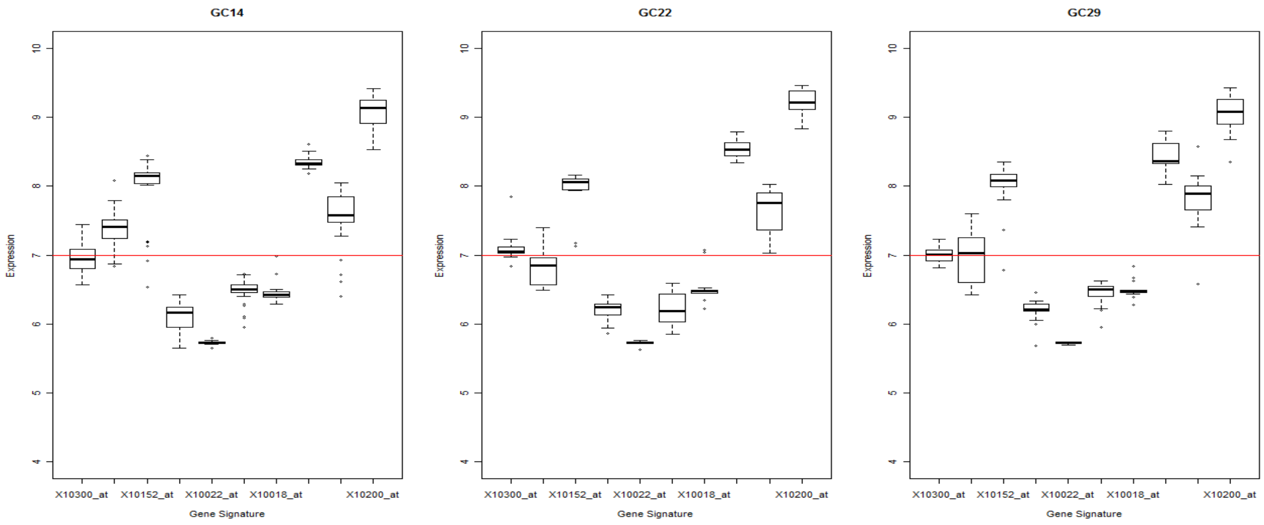


Figure 2: Expression levels of the three clusters gene signature.

Finally, 2000 bootsraps were performed with the gene signature in table 7 above with LDA as the classification function each time validated with the bootsraap test set. The mean misclasification error was 0.3228 which is approximately 32%. This was suggested to be high because of the similarity between clusters GC22 and GC29 that frequently leads to misclassification.

3.2.2. Two clusters analysis

Because of the high similarity between clusters GC22 and GC29, they were merged into one cluster and in a similar manner as in the three cluster analysis, features selection was performed as described in the methodology for two clusters settings. Table 8 below gives mean misclassification errors for all the various combinations of selection methods, top k genes and the classification functions. It can clearly be seen that the combination of top 10 genes, moderated t test as selection method and LDA as the classification function has the smallest misclassification error of 0.0807 with a standard error of 0.0356.

Table 8: Two clusters mean misclassification error rates and their standard errors in brackets.

Top	Selection Method	SVM	LDA	TBB	KNN	RF
5	t test	0.1078(0.0474)	0.0850(0.0507)	0.1519(0.0552)	0.1076(0.0432)	0.1130(0.0517)
	Wilcoxon test	0.1372(0.0332)	0.1553(0.0317)	0.1681(0.0486)	0.1569(0.0577)	0.1559(0.0317)
	Moderated t test	0.1279(0.0452)	0.0928(0.0354)	0.1654(0.0496)	0.1092(0.0379)	0.1254(0.0455)
10	t test	0.1106(0.0490)	0.0908(0.0452)	0.1561(0.0556)	0.1230(0.0398)	0.1093(0.0392)
	Wilcoxon test	0.1553(0.0486)	0.1586(0.0528)	0.2000(0.0637)	0.1619(0.0543)	0.1679(0.0532)
	Moderated t test	0.1101(0.0470)	**0.0807(0.0356)**	0.1535(0.0650)	0.1201(0.0522)	0.1134(0.0540)
20	t test	0.0897(0.0339)	0.1239(0.0390)	0.1446(0.0687)	0.1303(0.0460)	0.1078(0.0489)
	Wilcoxon test	0.1710(0.0524)	0.2148(0.0572)	0.2100(0.0670)	0.1646(0.0576)	0.1679(0.0453)
	Moderated t test	0.1078(0.0474)	0.1415(0.0467)	0.1504(0.0618)	0.1295(0.0544)	0.1120(0.0580)

Figure 3 is a graphical representation of the mean misclassification errors as a function of the top k genes and the selection methods. Based on this plot and the values of the mean misclassification errors the combination of top 10, moderate t test and LDA was selected because it is the curve with the lowest mean misclassification error.

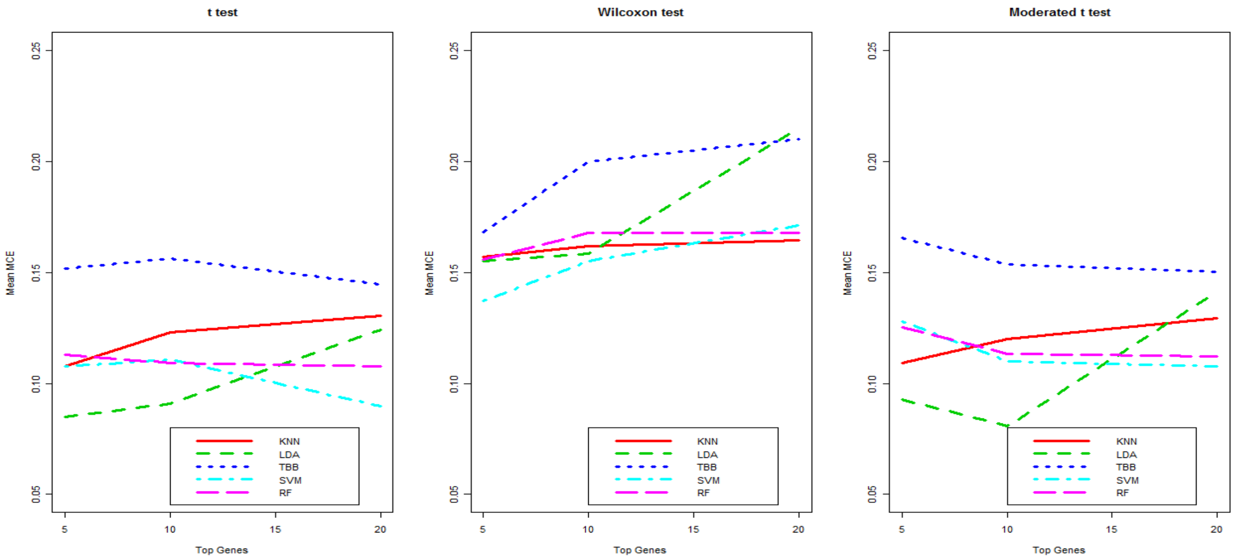


Figure 3: Mean misclassification error as a function of top k and selection method for 2 clusters.

For moderated t test as the selection method and top 10 genes to be selected, 2000 bootstraps were performed for gene selection and table 9 below gives the top 10 genes that were selected most of the times with their respective frequencies.

Table 9: Gene signature for the two clusters

Gene	10221_at	100287237_at	100141515_at	10095_at	100131187_at	100128071_at	100286909_at	100287098_at	10217_at	100129637_at
Freq.	2000	1926	1854	1582	1408	958	867	805	697	570

For this gene signature, the expression levels for each gene were compared by a box plot as shown in figure 4 below. From this figure, one clearly sees that the expression levels of the genes differ across the clusters. A gene by gene visualisation of the box plot makes this clearer than a block comparison.

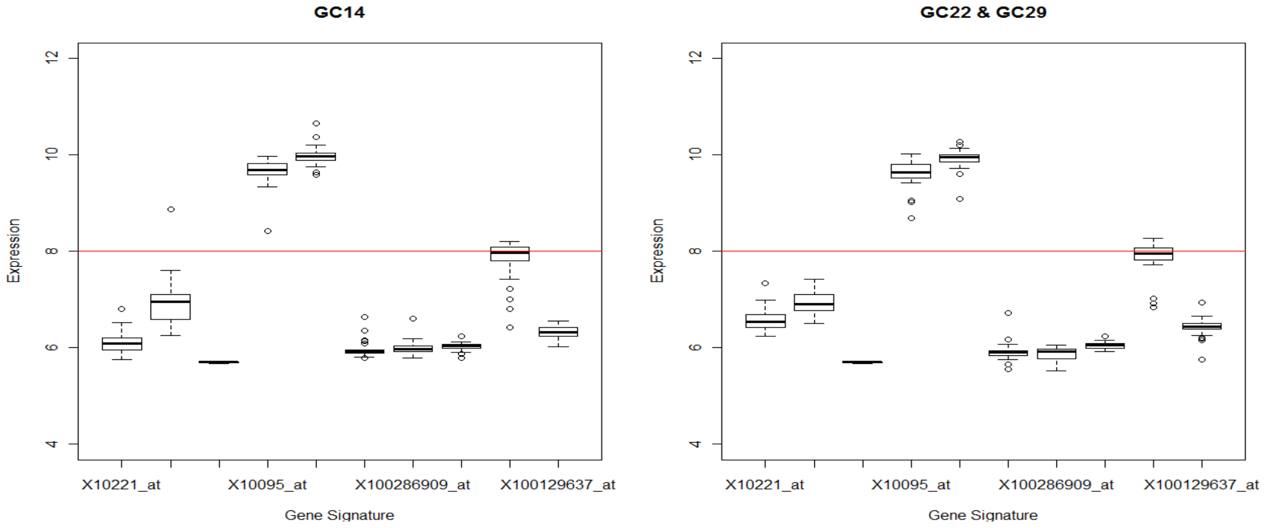


Figure 4: Expression levels of the two clusters gene signature.

From 2000 bootsraps performed with the gene signature in table 9 above and LDA as the classification function, the following results were obatined when validated with the bootstrap test sets; a mean misclassification error rate of 0.0724 which is approximately 7%, a sensitivity of 89% and a specificity of 96%.

4. DISCUSSIONS AND CONCLUSION

From the analysis of differentially expressed genes, parametric, nonparametric and resampling-based methods have shown that the three clusters of chemical compounds are not the same as reflected by the number of rejected hypotheses observed in this analysis. Though the three clusters were found to be dissimilar, it has been observed from the analysis of contrasts that clusters GC22 and GC29 are similar to each other as reflected by the no or few number of rejected hypotheses at 5% significance level. Clearly, assumptions are made about the distribution and/or variability across genes when using parametric tests like t test, moderated t test, F test and moderated F test. These assumptions may be violated leading to false results. Also, nonparametric tests like the Wilcoxon and the Kruskal-Wallis tests replace data with their ranks thereby leading to loss of information but have the advantage that the statistics are insensitive to measurement errors. Resampling-based (permutation) methods make no distributional assumptions about the statistic. The choice of an optimal method is not of much concern in this analysis and is left onto to the individual to choose which method best suits his/her desire since the goal was to determine using several possible methods if truly the three clusters are different. That notwithstanding, one clearly sees from Figures A1 (a) and (b) in the appendix, that there are quite a number of genes with large test statistics due to small within cluster (sample) variance. Hence, for this data, I will recommend resampling-based methods for the analysis of differential expression.

For classification analysis, seven classification functions were chosen and five of these functions were used at the early stage of the classifier building process. The choices of the classification functions are subjective, that is there is no standard motivation for these choices. They can equally be replaced by any of the numerous classification functions that exist in the literature. Also, for the choice of classification function made, it might be required to carry out parameters tuning to determine the optimal values for the parameters to be used. For instance, for support vector machines (SVM) it might be required to fine tune the optimal value for the cost parameter while for k nearest neighbour (KNN) one also needs to determine the optimal value for k. This is required because classifiers might perform slightly better with optimal values of their parameters than with default values though in some cases, the difference may not be noticeable. For this analysis, optimal value(s) for the parameter(s) was (were) not performed. Thus, it could be that if performed, the classifier with the lowest misclassification error might be different from the current classifier in each of the clusters (three or two) settings.

With respect to the classification analysis, it was observed that with the gene signature in Table 7 and linear discriminant analysis (LDA) as the classification function, the misclassification rate stood at approximately 32% for the three clusters setting. This was suspected to be high because of the close similarity between clusters GC22 and cluster GC29. Worth to mention is the fact that the LDA classification function could have been replaced by the feed-forward neural networks (NNET) classification function since they both had same misclassification errors at different values of top selected genes. After merging clusters GC22 and GC29 to one cluster and performing the classifier building process with the two formed clusters, the gene signature in Table 9 and still linear discriminant analysis (LDA) as the classification function confirms the assertion that the misclassification error could have been high in the three cluster setting certainly because of the similarity between clusters GC22 and GC29 in that the misclassification rate reduced drastically to approximately 7%. This misclassification rate seems to be very encouraging as also illustrated by high values of sensitivity and specificity obtained from a sensitivity-specificity analysis. This analysis revealed a sensitivity of 89% and a specificity of 96%.

In conclusion, I will recommend that if the goal of the experiment is to classify and predict the class of a chemical compound using gene expression data, then the researcher(s) should consider merging clusters GC22 and GC29 into one main cluster. This is because it has clearly been observed that based on the gene expression data, these clusters are not different and also, better classification and prediction results are obtained when these two clusters are merged and considered as one than when they are considered as separate clusters.

5. Software

The sofware used for the entire analysis was the statistical sofware R version 2.13.0 with its accompanying bioconductor packages as described in each subsection below.

5.1. Differential Expression

For differentially expressed genes analysis, the packages Biobase for base functions for bioconductor, limma for linear models for microarray data, samr for significance analysis of microarrays, multtest for multiple hypothesis testing and xlsx for writing to spreadsheets were utilised.

Differentially expressed genes analysis for the three clusters setting was performed using the F test, moderated F test, Kruskal-Wallis test and SAM. For F test, a linear model was fitted for the expression data of this gene against the vector of class labels using the method lm and this model was tested for no cluster effect using the method anova. The F statistic and the p-value for each gene were retained. In a similar manner, the Kruskal-Wallis test was performed by fitting a Kruskal-Wallis model for each gene using the method kruskal.test and the Kruskal-Wallis statistic and p-value also retained. In both cases, I controlled for multiple testing as described in the methodology using the package multtest. First the method mt.rawp2adjp was used to convert the raw p-values retained from the fits to adjusted p-values using the BH adjustment procedure. Then the raw p-values and adjusted p-values were then combined into a data frame. Secondly, for a vector of values of false discovery rate (FDR) to be controlled, the method mt.reject was used to return the number of genes rejected at each significance level with the return object *r* and the genes that were rejected with the return object *which*.

For moderated F test, the package limma was used. Firstly, an *"assaymat"* was created by converting the data into a matrix using the method as.matrix. Secondly, an expression set "*myexpset*" was created from the *"assaymat"* matrix using the method new. Next the design matrix was then created from the class labels using the method model.matrix and the model between the expression set and the design matrix was then fitted using the method lmFit and finally the empirical Bayes statistics and smoothed standard errors were computed for this fit using the method eBayes. The method topTable was then applied to this empirical Bayes fit with no multiplicity adjustment and with the BH multiplicity adjustment procedure at various significance levels to get the number and consequently the rejected hypotheses.

Comparison with significance analysis of microarrays (SAM) was done using the package samr. Firstly, a list object containing the expression data *X*, the class labels of the samples *Y*, the gene IDs, and the gene names is created then a SAM object is created using the method samr with arguments *resp.type="Multiclass", nperms=100;* corresponding to the response type and the number of permutations to be carried out. Secondly, the method samr.compute.delta.table was applied to the SAM object to produce a series of delta values and the delta value whose 90 percentile FDR corresponded to approximately 5% was chosen. With this delta value, the method samr.compute.siggenes.table was applied to the SAM object, the list object and the delta table to return a list of rejected hypotheses for this delta value. Finally, the method samr.plot was applied on the SAM object and the chosen delta value to produce the SAM plot showing the up and/or down regulated genes.

For the contrasts analysis, for each and every gene, each cluster was compared with the other using the t test via the method t.test, Wilcoxon test via the method wilcox.test and in a similar manner; the raw p-values retained from these comparisons were adjusted for multiplicity using the multtest package as described above. Also, the comparison made using the moderated t test was performed in the limma package by creating a contrast matrix from the design matrix using the method makeContrasts, then a contrast model was fitted using the method contrasts.fit and finally, the empirical Bayes statistics were computed from this fit by applying the method eBayes and in a similar manner, the method topTable was applied to the empirical Bayes contrast fit to get the number of rejected hypotheses and definitely the hypotheses at different significance level without multiplicity adjustment and with BH multiplicity adjustment. Comparison with the SAM procedure is exactly the same as in the three clusters settings except that the expression set and the class labels in this case were those of the clusters to be compared and the method samr now had the argument *resp.type="Two class unpaired".*

Lastly, from the statistics produced from the F test and t test, their computed standard errors and their p-values, a series of fold-change plots we produced using normal plots methods in R. Also, rejected hypotheses of one method we compared with those of the other methods using the operation *%in%*. And all through the analysis, outputs intended to be saved as a spreadsheet were done so using the method write.xlsx from the xlsx package.

5.2. Classification

Classification was performed using the package CMA and other required libraries include: Biobase for base functions for bioconductor, gbm for tree-based boosting, randomForest for random forest, limma for linear models for microarray data, class for k-nearest neighbours, MASS for linear discriminant analysis, nnet for feed-forward neural network, e1071 for support vector machines and xlsx for writing to spreadsheets.

Two functions **bestcombA** and **bestcombB** were built for both the three clusters and two clusters settings analyses each taking values *topk* and *method* corresponding to the top genes to be selected and selection method respectively. Each also returns a list object containing the mean misclassification error and standard error of the mean for each combination of top genes, selection method and classifier. Within each function, a “for loop” from 1 to 1000 was included containing the CMA methods: GenerateLearningsets, GeneSelection, classification and compare. For the method GenerateLearningsets, the vector of class labels *Y* is used together with *method="MCCV"* corresponding to Monte Carlo Cross Validation, *niter=2*, corresponding to two iterations thus making a total of 1000x2 generated learning and test sets, *ntrain= floor(2/3*length(Y))* corresponding to the number of samples to be in each learning set and *strat=TRUE* to take into account the number of samples per class label in both the learning and the test set.

For each learning set, genes were selected using the method GeneSelection with arguments *X* the expression matrix and *method* taking the value provided in **bestcombA** or **bestcombB** and for each batch of selected genes, five classifiers (of *dldaCMA, ldaCM, gbmCMA, knnCMA, nnetCMA, rfCMA, svmCMA* corresponding to diagonal linear discriminant analysis, linear discriminant analysis, tree-based boosting, k-nearest neighbours, feed-forward neural networks, random forest and support vector machines respectively) were built using the method classification and taking amongst its arguments *X, Y* and *topk* genes provided as argument to **bestcombA** or **bestcombB.** Finally, the results of the five classifiers were joined as a list object using the method join and compared using the method compare and the misclassification error for each classifier was retained. This was repeated till the "for loop" was terminated and the mean misclassification error and standard error of the mean was computed for each classifier. The functions **bestcombA** and **bestcombB** were then called several times with different combinations of *topk* (k= *5,10* or *20*) and *method* [one of *f.test*, *kruskal.test* or *limma for* **bestcombA** and one of *t.test*, w*ilcox.test,* limma for **bestcombB**]. Values returned from these several calls were then combined, formatted and saved as in table 6 or 8 and also plotted as in figure 1 or 3.

In each data setting (three or two clusters), for the selected top k genes and selection method with the smallest misclassification error and for 2000 bootstraps, the sample was again split into learning and test sets using the method GenerateLearningsets with same arguments as described above and for each learning set, genes were selected using the method GeneSelection with the argument method taking the value of the chosen method. For each selected batch of genes, the top k (the chosen value) genes were retained using the method toplist. At the end of the 2000 bootstraps the retained genes were formatted as a table using the method table and ordered in descending order of their frequency counts. The top k genes in this table were returned as the gene signature. Based on this signature, the gene expression matrix *X* was reduced to only the gene expression data of the signature say *X** by eliminating rows of genes not belonging to the signature. With the gene signature a gene by gene box plot for all the clusters was plotted using the method boxplot and taking into account which samples corresponds to which cluster as shown in figures 2 and/or 4.

Finally, with the chosen classifier and for 2000 bootstrap, learning and test sets were generated in a similar manner as described above using the mehod GenerateLearningsets. For each learning set, the classifier was built using *X*, Y.* The performace of the classifer was then evaluated using the method evaluation taking argument *measure* = "misclassification", "sensitivity" or "specificity".

6. REFERENCES

I. Agresti, A. & Finlay, B. (1997). Statistical Methods for the Social Sciences. (3rd Ed.). Prentice Hall, New Jersey. USA.

II. Breiman, L. (2001). Random Forest. *Machine Learning,* **45**:5-32

III. Conover, W. J. (1998). Practical Nonparametric Statistics. (3rd Ed.) . John Wiley & Sons, New York. USA.

IV. Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistic*, **29**:1189-1232

V. Han, J. & Kamber, M. (2006). Data Mining Concepts and Techniques (2nd Ed.). Elsevier Inc., San Francisco, USA.

VI. Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning (2nd Ed.). Springer, New York, USA.

VII. Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, William. (2005). Applied Linear Statistics Models (5th Ed.). McGraw-Hill, New York. USA.

VIII. Lehmann, E. L. (2006). Nonparametrics Statistical Methods Based on Ranks (1st Ed.). Springer, New York. USA.

IX. Lin, D., et al (2008). An Investigation on Performance of Significance Analysis of Microarray (SAM) for the Comparisons of Several Treatments with one Control in the Presence of Small-variance Genes. *Biometrical Journal,* **50**:5, 801–823

X. Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D. & Bijnens, L. (2010). Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R. Springer.

XI. McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons, New Jersey. USA

XII. Mendenhall, W. & Sincich, T. (2007). Statistics for Engineering and the Sciences. Pearson, New Jersey. USA.

XIII. Meyer, D. (2011). Support Vector Machines. *R-News*, **1:**3.

XIV. Pollard, K. S., Dudoit, S. & Van der Laan, M. J. (2004). Multiple Testing Procedures: R multtest Package and Applications to Genomics. *U.C. Berkeley Division of Biostatistics Working Paper Series.* Working Paper 164. http://www.bepress.com/ucbbiostat/paper164.

XV. Ripley, B. (1996). Pattern Recognition and Neural Networks. *Cambridge University Press.*

XVI. Scholkopf, B. & Smola, A. (2002). Learning with kernels. MIT Press, Cambridge, MA, USA.

XVII. Slawski, M., Daumer, M. & Boulesteix, A. (2008). CMA-a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*, **9**:439. (http://www.biomedcentral.com/1471-2105/9/439).

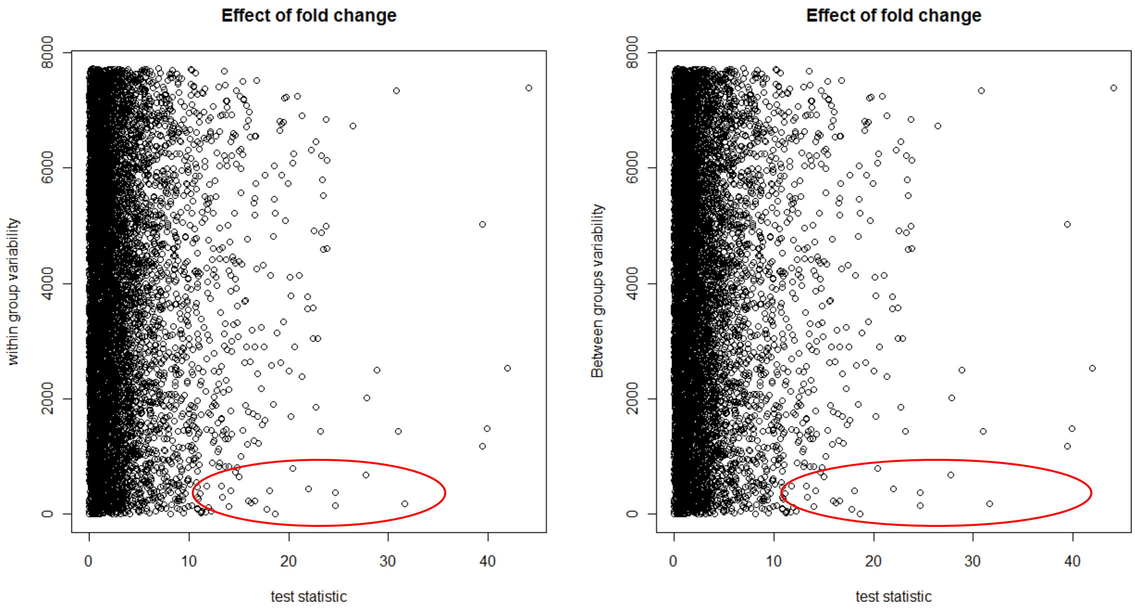
XVIII. Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**:1, 3. (http://www.bepress.com/sagmb/vol3/iss1/art3).

XIX. Tan, P., Steinbach, M. & Kumar, V. (2005). Introduction to Data Mining. Pearson, New York, USA.

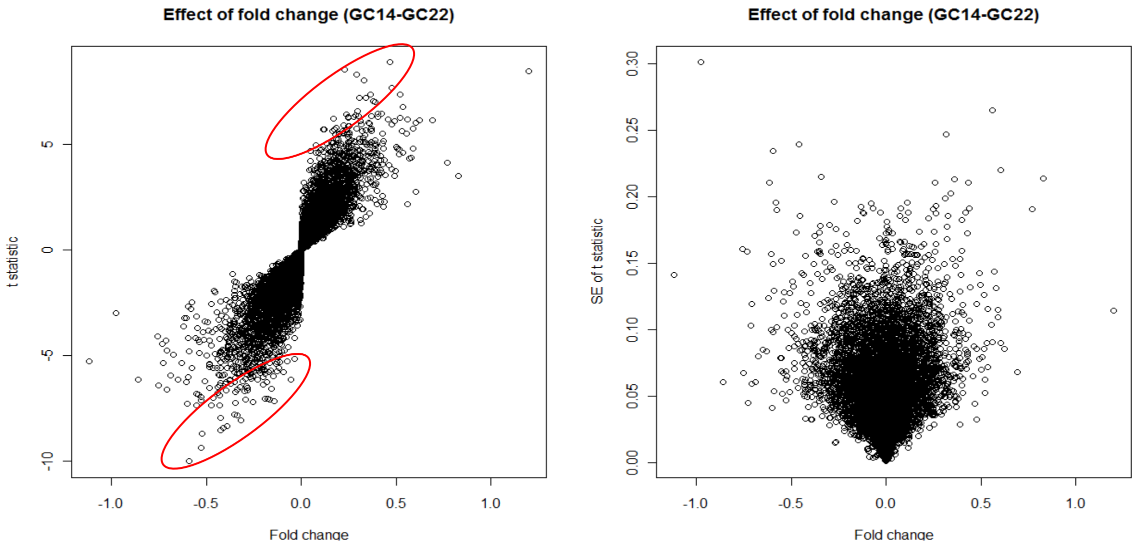
XX. Tibshirani, R., Chu, G., Narasimhan, B. & Li, J. (2011). Significance Analysis of Microarrays. http://www-stat.stanford.edu/~tibs/SAM

7. APPENDIX

Figure A1: Effects of fold change and small variance genes

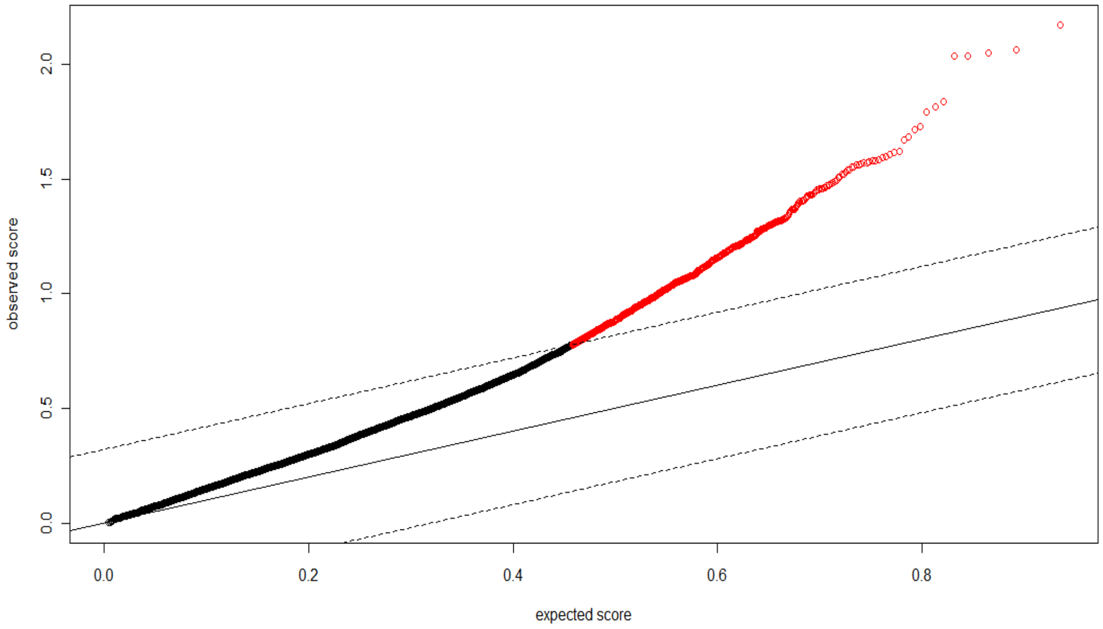


a. Effect of fold change from the F statistic

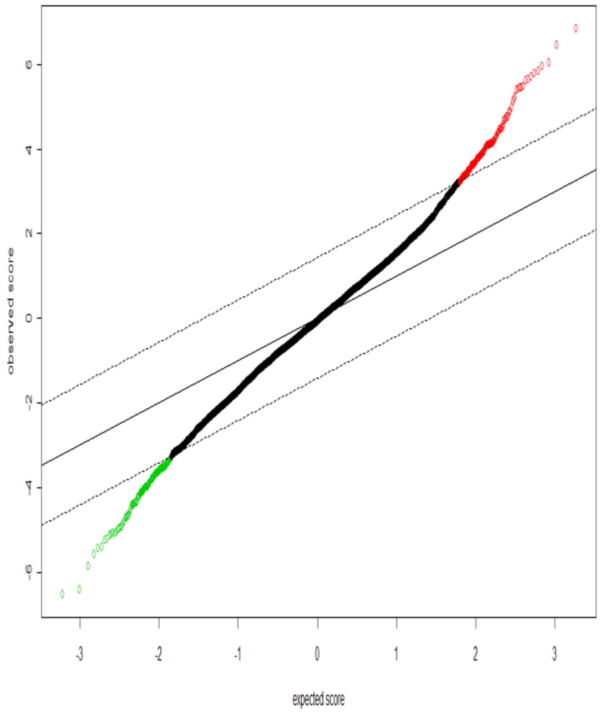


b. Effect of fold change from the F statistic

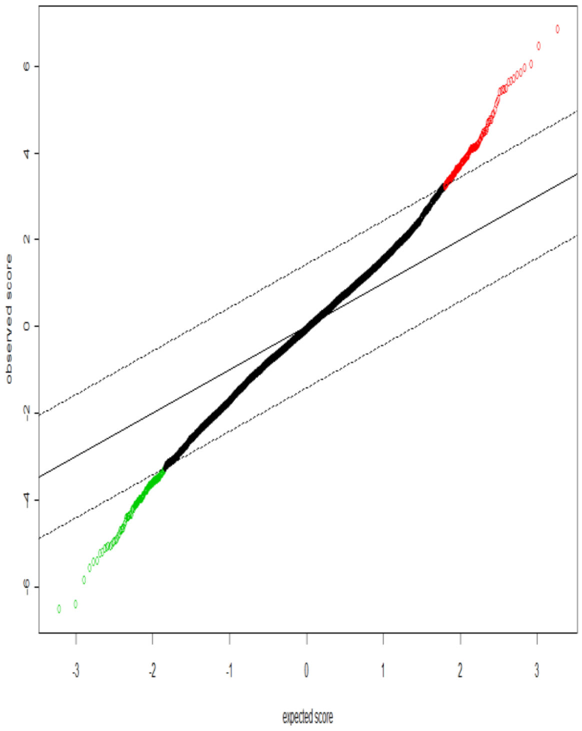
Figure A2: SAM plots.



a. SAM F statistic (delta = 0.30)



b(i) GC14-GC22; SAM t statistics (delta= 1.43)



b(ii) GC14-GC29; SAM t statistics (delta= 1.30)

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Classification and class prediction for different chemical structures using gene expression data

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt
behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -,
vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten
verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Jong, Victor Lih

Datum: **12/09/2011**