Arrow Plot: a new tool for the analysis of gene expression profiles

Carina Silva-Fortes*[1], Maria Antónia Amaral Turkman[2] and Lisete Sousa[2]

[1]Natural and Exact Sciences Department, Higher School of Health Technology of Lisbon of Polytechnic Institute of Lisbon and Center of Statistics and Applications of University of Lisbon, Lisbon, Portugal

[2] Department of Statistics and Operational Research, Faculty of Sciences of University of Lisbon, and Center of Statistics and Applications of University of Lisbon, Lisbon, Portugal

Email: Carina Silva-Fortes*- carina.silva@estesl.ipl.pt; Maria Antónia Amaral Turkman - antonia.turkman@fc.ul.pt; Lisete Sousa - lmsousa@fc.ul.pt;

*Corresponding author

Abstract

Background: A common task in analyzing microarray data is to determine which genes are differentially expressed across two (or more) kinds of tissue samples or samples submitted under experimental conditions. Several statistical methods have been proposed to accomplish this goal, generally based on measures of distance between classes. It is well known that biological samples are heterogeneous because of factors such as molecular subtypes or genetic background that are often unknown to the experimenter. For instance, in experiments which involve molecular classification of tumors it is important to identify significant subtypes of cancer. Bimodal or multimodal distributions often reflect the presence of subsamples mixtures. Consequently, there can be genes differentially expressed on sample subgroups which are missed if usual statistical approaches are used. In this paper we propose a new graphical tool which not only identifies genes with up and down regulations, as well as genes with differential expression in different subclasses, that are usually missed if current statistical methods are used. This tool is based on two measures of distance between samples, namely the overlapping coefficient (OVL) between two densities and the area under the receiver operating characteristic (ROC) curve. The methodology proposed here was implemented in the open-source R software. **Results:** This method was applied both to a publicly available data set and to a simulated data set. We compared our results with the ones obtained using some of the standard methods for detecting differentially expressed genes, namely Welch t-statistic, fold change (FC), rank products (RP), average difference (AD), weighted average difference (WAD),

1

moderated t-statistic (modT), intensity-based moderated t-statistic (ibmT), significance analysis of microarrays (samT) and area under the ROC curve (AUC). On both data sets all differentially expressed genes with bimodal or multimodal distributions were not selected by all standard selection procedures. We also compared our results with (i) area between ROC curve and rising area (ABCR) and (ii) the test for not proper ROC curves (TNRC). We found our methodology more comprehensive, because it detects both bimodal and multimodal distributions and different variances can be considered on both samples. Another advantage of our method is that we can analyze graphically the behavior of different kinds of differentially expressed genes. **Conclusion:** Our results indicate that the *arrow plot* represents a new flexible and useful tool for the analysis of gene expression profiles from microarrays.

Background

Genome-wide expression analysis is an increasingly important tool for identifying gene function, disease-related genes and transcriptional patterns related to drug treatments. Microarrays enable the simultaneous measurement of the expression levels of tens of thousands of genes and have found widespread application in biological and biomedical research. Increasing numbers of multi-class microarray studies are performed, but the vast majority continues to be two class (binary) studies, for example when both control and a treatment are examined [1–4]. The objective of the study in most of them, is to determine the genes that are differentially expressed between the two classes. Differentially expressed genes are usually detected using statistics based on means or medians. However, if there are genes differentially expressed on different subclasses, those techniques do not select them because either mean or median values tend to be similar between the considered groups.

Genes with a bimodal or a multimodal distribution within a class (considering a binary study) may indicate the presence of unknown subclasses with different expression values [5,6], meaning that there are two separate peaks in the distribution: one peak due to a subclass clustered around a low expression level, and a second peak due to a subclass clustered around a higher expression level. As a consequence, the identification of such subclasses may provide useful insights about biological mechanisms underlying physiologic or pathologic conditions. In cancer research, a common approach for prioritizing cancer-related genes is to compare gene expression profiles between cancer and normal samples, selecting genes with consistently higher expression levels in cancer samples. Such an approach ignores tumor heterogeneity and is not suitable for finding cancer genes that are overexpressed in only a subgroup of a patient population. As a result, important genes differentially expressed in a subset of samples can be missed by gene selection

criteria based on the difference of sample means [7].

The particular application that motivated our work concerns the development of a methodology which could simultaneously identify up- and down-regulated genes and differentially expressed with bimodal or multimodal distributions with similar means on both groups. For convenience, the latter case is referred to as *special genes*.

Different statistical tests have been proposed to select differentially expressed genes [8–11]. Among them, is the receiver operating characteristic (ROC) analysis, which is widely used to evaluate a diagnostic system but can be interpreted as a measure of separation between two distributions.

A ROC curve displays the relationship between the proportion of true positive (sensitivity) and false positive (1-specificity) classifications resulting from each possible decision threshold value in a two class classification task [12]. These proportions depend on the classification rule and in general higher values of the marker are associated with the case group. However, if ROC analysis is blindly applied to select genes with differentially expressed, i.e., keeping the same classification rule for all genes in an experiment, not proper ROC curves (NPROC) [11] can be produced because genes with positive and negative regulation have opposite classification rules. NPROC curves are obtained when they cross or are below of the reference line (Figure 1 C-E).

Genes can be ranked using the area under the ROC curve (AUC) [10, 11], a common measure of discrimination which should ranged between 0.5 and 1, but for NPROC curves AUC can have values below 0.5.

Nevertheless, different scenarios can lead to NPROC curves, for instance, when the means of the two groups are similar and one of the groups has a bimodal distribution (Figure 1 C-D) (or multimodal), or when both distributions are unimodal with similar means and significant different variances (Figure 2). On both cases the corresponding ROC curve will have a sigmoidal-shape with an AUC close to 0.5.

Proper binormal model [13] and contaminated binormal model [14] are methods that force the ROC curves to be set above the reference line when they are not proper and consequently the AUC will be higher than 0.5. However in the context of this work, not proper ROC curves have an important role in the selection of different kinds of differentially expressed genes.

Since it is not possible to decide, beforehand, what should be the direction of the classification rule, we considered the same classification rule for all of the genes, i.e., values of expression levels above the threshold correspond to up-regulation. In that sense, AUC values near 1 will correspond to up-regulated genes, AUC values near 0 will correspond to down-regulated genes, and special genes (Figure C-D) will

have an AUC around 0.5. However, regardless the type of distributions, if means are similar (Figure 1 B, Figure 2), AUC will be near 0.5. So, using AUC is not sufficient to select special genes.

We used the overlapping coefficient (OVL) to further separate these different situations which produce values of AUC near 0.5. Bradley [15] and Inman and Bradley [16] promote the use of OVL as an intuitive measure of the similarity between two probability distributions. Graphically, OVL is the area where the densities of the two distributions overlap when plotted on the same axes.

We propose using AUC and OVL simultaneously to select different types of differentially expressed genes and plotting OVL against AUC we get a picture which we named as *arrow plot* (Figure 5).

If we consider that groups have different variances, special genes can be mixed with genes which are not differentially expressed as illustrated on Figure 2, that is, genes with unimodal densities, with similar means but significantly different variances. These genes will have AUC values around 0.5 and low OVL values. We developed an algorithm based on finding bimodality (or multimodality) using kernel densities estimates, to identify genes under those conditions allowing their separation from the special genes.

Nonparametric techniques are used to estimate AUC and OVL. To estimate AUC we used the Mann-Whitney U statistic [17], which is equivalent to the trapezoidal rule for integration. For the OVL we developed an algorithm where a naive kernel density estimator [18] is used to develop a nonparametric estimator of OVL.

We first describe the algorithm and then the performance of our method is evaluated by comparing the gene expression profiles in two different classes using data from a publicly data set [6] and from a simulated data set. The first data set consists of 14 different samples of normal circulating B cells (controls) and 20 heterogeneous lymphomas (experimental group) [6]. The gene expression data were obtained on 4026 genes. The simulated data set consists of 10000 genes generated from a lognormal distribution, where each group sample has 30 arrays. We compared our results obtained for publicly data with the results obtained by Parodi et al. [11], where area between ROC curve and rising area (ABCR) and test for not-proper ROC curves (TNRC) were used. On both data sets we compared the results with some usual methods to assess the relative performance of different statistical gene ranking measures.

All the analysis was performed using the open-source R software [19] and packages from Bioconductor [20].

Results and Discussion
Algorithm description

For illustrative purposes, we divided the algorithm in two parts (algorithm 1 and algorithm 2). The first one describes the OVL estimation (Figure 3) and the second describes the selection of different kinds of differentially expressed genes (Figure 4).

The OVL estimation was based on a non-parametric form with densities estimated using kernel functions. Figure 3 shows the pseudo-code which implements the OVL estimation and tables 1 and 2 describe the notation and functions used on it. The OVL values are computed by finding the points that belong to the area of intersection of the two densities (Figure 3: lines 1-21) and the jump points between densities, which are estimated by interpolation (Figure 3: lines 24-44). The points are combined into one set and sorted in ascending order (Figure 3: line 50). Finally OVL is estimated using a trapezoidal rule considering a non-uniform grid-spacing (Figure 3: line 51).

The selection of differentially expressed genes is based on simultaneous analysis of OVL and AUC. The *arrow plot* is obtained by plotting OVL on abscissas and AUC on ordinates. Figure 4 shows the pseudo-code which implements the algorithm to select differentially expressed genes based on these two measures and tables 3 and 4 present the notation and functions used on it.

Selection of differentially expressed genes with positive regulation (Figure 4: line 6-7) and negative regulation (Figure 4: line 9-10), is made according to arbitrarily selected cutoff points for AUC and OVL. However, AUC values are expected to be close to 1 for up-regulated and close to 0 for down-regulated genes and OVL will have low values on both situations. Selection of special genes is performed in two steps. The first step consists on the selection of genes with AUC values near 0.5 and low values of OVL (Figure 4: lines 12-13). Since the variances on both groups can be different, it is possible to find genes with no-differential expression mixed with the special ones. Accordingly, the second step aims at removing the genes without differential expression, through the bimodality analysis.

Bimodality (or multimodality) is analyzed based on the behavior of the ordinates of the kernel based estimated densities of both groups, considering only the gene list that is selected in the first step mentioned above (Figure 4: line 13). The points of both densities are ordered in increasing order of abscissas (Figure 4: lines 22-23). If an ordinate is equal or less than the ordinate immediately after, it is assigned a label 1 and 0 otherwise (Figure 4: lines 25-28 and 38-41). This allows us to analyze the variation of the density over the observed range. Considering only the points where the function is increasing, if the differences between the ranks of adjacent ordinates is 1, the distribution is expected to be unimodal, otherwise the

5

distribution will be bimodal or multimodal (Figure 4: lines 30-33 and lines 43-46). To consider a gene to be special it is enough to find bimodality in one of the groups (Figure 4: lines 50-54), yet it is of interest to analyze in which group bimodality is observed, and this is possible using different color labels on the *Arrow plot*.

Performance

The running time of the algorithm in a data set with 10000 genes, takes less than 60 minutes on a 533 MHz Pentium.

Lymphoma data

From a total of 4026 genes, our method selected 178 differentially expressed genes, where 68 corresponded to up-regulated genes, 90 to down-regulated and 20 corresponded to special genes. We used AUC$\geq$0.9 and OVL<0.5 to select up-regulated genes, AUC$\leq$0.1 and OVL<0.5 to select down-regulated genes and OVL<0.5 and 0.4<AUC<0.6 to select special genes. Thresholds were chosen arbitrarily, although an analysis of the the *arrow plot* (Figure 5) could help on deciding which thresholds to use.

Table 5 shows the 20 selected special genes. Genes are listed in ascending order of OVL, which ranged between 0.389 and 0.499. AUC values ranged between 0.407 and 0.593. Bimodality was tested on the 20 special genes; 2 genes have bimodality in the control group, 5 genes on the experimental group and 13 genes on both. For the 20 special genes, kernel densities and their corresponding empirical ROC curves can be analyzed in Figure 6. All the selected genes had a sigmoidal-shaped ROC curve.

Among the 20 special genes selected list (table 5), 3 have an unknown regulatory function. All the remaining 17 genes are related with proteins encoding. GENE3323X (*BCL7A*) and the GENE3388X (*Immunoglobulin J chain*) are presented in other clones in the same data set, GENE3322X and GENE3389X respectively. Alizadeh et al. [6] noted that *BCL7A* gene can be altered by translocations in lymphoid malignancies. Table 6 shows the biological properties of the selected genes.

We compared our results with the ones obtained by Parodi et al. [5], where ABCR and TNRC statistics were used. According to the highest TNRC value, they selected 16 differentially expressed genes, from a total of 1607 differentially expressed genes. Eight of them are considered to be special according to our methodology. The remaining 8 genes of their list have AUC and OVL values slightly higher than the considered cutoff points on our study. However, if we choose threshold values for AUC and OVL to catch those genes, we will select 85 more special genes.

6

Nine feature selection methods were applied to the full data set, namely Welch t-statistic, fold change (FC), rank products (RP) [21], average difference (AD) [22], weighted average difference (WAD) [23], moderated t-statistic (modT) [24], intensity-based moderated t-statistic (ibmT) [25], significance analysis of microarrays (samT) [26] and area under the ROC curve (AUC). We assessed the overlap between gene lists produced by different feature selection methods and ranked lists of differentially expressed genes were produced. We examined the top 20 mostly highly ranked genes, and for all methods the 20 special genes selected by our methodology are missed.

Simulated data

Ten thousand genes were simulated (see Methods for details), where 9500 are non-differentially expressed, and 500 are differentially expressed where: 200 are up-regulated, 200 are down-regulated and 100 are special. Analyzing the *arrow plot* (Figure 7), we considered OVL< 0.3 and the following thresholds for AUC: to select up-regulated genes AUC≥ 0.9; to select down-regulated genes AUC≤ 0.1; and to select special genes $0.4 < AUC < 0.6$. In the *arrow plot* we can observe the distribution of the truly 500 differentially expressed genes, and we can conclude that 95% of them were selected by our methodology. In the first step of the algorithm to select special genes (Figure 4), 33 special genes candidates were selected. Through the second step we found that all of the genes had bimodality in at least one of the groups. We can conclude that our algorithm to detect bimodality performed with 100% of accuracy on that list. We also compared the 33 special gene list selected by our method with the 50 top mostly highly ranked genes by the methods discussed in the earlier example, and only the Fold Change method selected 10 genes from the simulated special gene list.

Conclusions

We have presented a graphical and computational method for microarray experiments which allow the identification of genes that express differently under two conditions even if the behavior in average is similar. The main objective of this work was to select differentially expressed genes due to the presence of different subclasses, which could give important information about their inherent biological functions, and that are usually missed by usual methods.

AUC and OVL statistics were used to achieve this goal. Both statistics are invariant when a suitable common transformation is made on variables [12, 16], and on microarray data analysis log transformations are widely used. *Arrow plot* is obtained by plotting OVL against AUC. This plot is easily interpreted

because both statistics range between 0 and 1, and in addition to detecting genes with up- or down-regulation, *arrow plot* is also able to detect special genes, however for the latter genes a bimodality analysis needs to be added.

Non-parametric techniques were used because they eliminate the need to specify parametric models. The non-parametric kernel density method has few assumptions about the form of the distributions. This is attractive because it can be used on thousands of genes on an automatic way. The disadvantage of non-parametric techniques is that it results in a loss of efficiency. Yet, the loss of efficiency is balanced by the reduction of the risk of misinterpreting the results by incorrectly specifying a parametric form for the distribution.

The proposed algorithm is especially useful in situations where bimodality exists in the gene expression data. The proposed methodology outperforms other well known methods for detecting different kinds of differentially expressed genes. Future work includes further evaluation of this methodology on other real data sets.

Arrow plot is an exploratory graphical tool for microarray experiments, useful in identification of different kinds of differentially expressed genes, particularly genes with a special behavior which are not detected by usual methods and yet can bring relevant biological information. This methodology can be used in all platforms.

Methods
Data sets
Lymphoma Data

We used microarray data provided by the study of Alyzadeh et al. (2000) [6] publicly available at the website http://llmpp.nih.gov/lymphoma/data/figure1/. They used a special microarray called Lymphochip, where they selected genes that are preferentially expressed in lymphoid cells and genes with known or suspected roles in important processes in immunology or cancer. They used these microarrays to characterize gene expression patterns in the three most prevalent adult lymphoid malignancies: DLBCL (diffuse large B-cell lymphoma), FL (follicular lymphomas) and CLL (chronic lymphocytic leukemia). They also profiled gene expression in purified normal lymphocyte subpopulations under a range of activation conditions (see original paper for more details [6]). Fluorescent cDNA probes, labelled with the Cy5 dye, were prepared from each experimental messenger RNA sample. A reference cDNA probe, labelled with the Cy3 dye, was prepared from a pool of mRNAs isolated from nine different lymphoma cell lines.

Each Cy5-labelled experimental cDNA probe was combined with the Cy3-labelled reference probe and the mixture was hybridized to the microarray. The fluorescence ratio was quantified for each gene and reflected the relative abundance of the gene in each experimental mRNA sample compared with the reference mRNA pool. The ratio values were log transformed (base 2) and stored in a table (rows, individual cDNA clones; columns, single mRNA samples). The data set that we used in our study is part of the original one, and was the same used in the study of Parodi et al. [5]. This database included 4026 gene expression profiles, where the control group had 14 samples of normal circulating B cells (NBC), where 6 were highly stimulated and 8 slightly or not stimulated samples. The experimental group had 20 heterogeneous lymphomas by pooling 9 samples of FL and 11 samples of CLL. Both classes included two subclasses, namely: 6 heavily stimulated and 8 slightly stimulated or unstimulated samples in controls and 9 follicular lymphomas and 11 chronic lymphocytic leukemias in experimental group. Parodi et al. [5] estimated missing data by the method proposed by Troyanskaya et al. [22].

Simulated data

We conducted a simulation study in order to evaluate the detection performance of the proposed method. Most studies of microarray data assumed normality assumptions. However, there is relatively little literature on evaluating the normality of the data. Part of the problem is that most microarray data sets include large amounts of biological variability and/or small sample sizes. Biological variability makes it difficult to determine the source of the non-normality (non-normal data sets could simply be mixtures of normal data sets). Small samples do not have the power to be able to make claims about the distribution of the data.

It is well known that raw microarray data (across all platforms) are highly skewed (usually skewed right) with many extreme values, so, simulated data sets were generated by drawing case and control samples from lognormal distributions, and log transformation is used to offset the skewness. Consider X a random variable which represents the expression levels in the control sample, were $X \sim logN(\mu_x, \sigma_x)$ and Y a random variable which represents the expression levels in case sample, were $Y \sim logN(\mu_y, \sigma_y)$.

For case and control samples we simulated $n_1 = n_2 = 30$ microarrays and a total of 10000 genes. This sampling was performed independently, even knowing that individual gene expression levels are far from being independent. In a typical microarray experiment, we expect to see a combination of non-differentially and differentially expressed genes (approximately 5% to 10% of the data). Hence, we simulated 500 genes differentially expressed and 9500 not differentially expressed. From the 500 differentially expressed genes,

200 were up-regulated, 200 were down-regulated and 100 corresponded to special genes.

Four characteristics of the data are considered in this simulation: mean (μ), variance (σ^2), the magnitude of difference between control and case samples and bimodality of the distributions. Hence, several combinations of these parameters were considered.

Wile simulating values for expression levels of genes not differentially expressed, we considered that the difference between the mean of the control and case arrays ranged between -0.9 and 0.9. To provide several patterns of density distributions we considered variances with differences ranging from 0 and 12.25. The effect of changing σ does not seem to affect these genes because all arrays came from the same nearly mean vector. However, some of these genes will be mixed with the special ones when the variances between case and control samples are significantly different.

Genes with up-regulation and down-regulation were generated considering the difference between the mean of the case and control arrays ranging from 3.5 to 13.5 for up-regulation, and -13.5 to -3.5 for down-regulation and the differences between the variances for both situations ranged from 0 to 12.25.

Gene expression distribution of a special gene was considered as a mixture of two lognormal distributions in one of the groups. If X is a random variable following this distribution, we write $X \sim \Theta_{\alpha,\mu_0,\sigma_0,\mu_1,\sigma_1}$ with the probability density function (PDF) $f(x) = \alpha \log N(x; \mu_0, \sigma_0) + (1 - \alpha) \log N(x; \mu_1, \sigma_1), x > 0$, where $\log N(x; \mu, \sigma)$ denotes the PDF of a lognormal distribution with location and scale parameters $\mu \in \Re$ and $\sigma > 0$, respectively, and $\alpha \in (0, 1)$ specifies the contribution to the total of the two single lognormal components. The parameters μ and σ become the mean and the standard deviation of the normal distribution upon log transformation of the lognormal random variable. To simulate special genes we considered bimodality in one of the groups. For the mixture we left $\mu_0 = 3.5$ unchanged and gradually increased μ_1 from 7 to 17, and left $\sigma_0 = \sigma_1 = 1.2$ unchaged. The α parameter ranged between 0.3 and 0.7. For the other group we considered an lognormal density with location parameter approximately equal to $\alpha \times \mu_0 + (1 - \alpha) \times \mu_1$.

Finally we took the logarithms of the 10000 expression levels on both groups to offset the skewness.

Non-parametric OVL

The overlapping coefficient refers to the area under two density functions simultaneously [15]. OVL is formally defined by (1):

$$OVL = \int_c \min[f_X(c), f_Y(c)]dc, \tag{1}$$

where f_X and f_Y are the density functions of the random variables X and Y respectively. The results are directly applicable to discrete distributions by replacing the integral with a summation.

The estimation of OVL was based on a non-parametric procedure with densities estimated using kernel functions. A kernel function $K(.)$ is defined as a continuous, limited and symmetric function, with the property that its indefinite integral is equal to unity, $\int K(u)du = 1$. The typical form of a kernel density estimator is given by (2):

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

where h is the bandwidth, known as the shaping parameter and $(x_1, ..., x_n)$ is the sampling vector.

For the purpose of this work, we chose as kernel function a standard normal distribution $(2\pi)^{-\frac{1}{2}} exp(-\frac{1}{2}u^2)$. More than the choice of the kernel function, the choice of the bandwidth, h, is what drives the kernel estimator. A choice of the bandwidth h satisfying some optimal criteria [27], is given by (3):

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-\frac{1}{5}}, \tag{3}$$

where σ is the standard deviation.

However this choice may tend to over-smooth the distribution if the population is multimodal. A better result may be obtained by using the interquartile range, R [28]. If the distribution of interest is bimodal, using interquartile range may over-smooth further. Therefore the use of an adaptive measure of spread is recommended (4):

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \min\left(\sigma, \frac{R}{1.34}\right) n^{-\frac{1}{5}}. \tag{4}$$

The function `density` from `R` uses as default settings the normal kernel and the value for h given in (4). In our calculations of the OVL, we used this function from `R`, with the default settings to estimate the densities, and for the trapezoidal rule needed to computed the area, we used the function `trapz` of the library `caTools`.

Non-parametric AUC

ROC curve assesses the effectiveness of a continuous diagnostic marker in distinguishing between two independent populations. In a standard situation a case is assessed positive if the corresponding marker value is greater than a given threshold value. Associated with any threshold value is the probability of a

true positive (sensitivity) and the probability of a true negative (specificity). Let X be the random variable for the marker on the control group and Y the random variable for the marker on the case group. For any given threshold value c, sensitivity is given by $P(Y > c) = 1 - F_Y(c)$, and specificity is given by $P(X \leq c) = F_X(c)$. The theoretical ROC curve is a function $ROC(t) = 1 - F_Y\left[F_X^{-1}(1-t)\right]$, where $t = 1 - F_X(c)$, that is 1-specificity. Hence, the ROC curve plots is 1-specificity against the sensitivity calculate for different values of the threshold c. The area under this curve (AUC) measures how well the marker discriminates the two groups involved and is given by $P(Y > X)$. Note that these definitions are a consequence of the assumption that high values of marker are associated with the case group.

The simplest non-parametric estimation method for the ROC curve involves using empirical cumulative distribution functions. The empirical cumulative distribution function is defined for any given value c, to be the observed percentage of sample values less than or equal to c. The resulting estimated ROC curve is an increasing step function on the unit square. The area under this curve is equal to the Mann-Whitney U-statistic and provides an unbiased non-parametric estimator for the AUC [17]. Bamber [29] showed that the AUC, when calculated using the trapezoidal rule, is equal to the MannWhitney U-statistic.

This method was performed using functions from the `ROC` library from Bioconductor.

Arrow Plot

Plotting OVL against AUC we obtain a graph which we called *arrow plot*. In order to identify different kinds of differentially expressed genes, it is necessary to select appropriate cutoff points both for the AUC and OVL. Differentially expressed genes will have low values for the OVL, say less than 0.5. Up-regulated genes will correspond to AUC near 1, down-regulated genes will correspond to AUC values near 0, and special genes will have AUC values around 0.5. An algorithm to check for bimodality is added, where special genes are highlighted using different colors depending whether bimodality is verified in case or control group.

TNRC and ABCR statistics

Parodi et al. [5] developed two new ROC based methods to identify differentially expressed genes that may correspond both to proper and to not proper ROC curves. TNRC (test) is a test to identify not proper ROC curves and ABCR (area below) statistic represents a measure of the distance between the distributions of gene expression in two classes.

The ABCR statistic is obtained using the empirical ROC curve, where ties are not considered. In that

sense, if n_0 is the number of individuals observed with X (considering the same notation as in non-parametric AUC section) and n_1 the number of individuals observed with Y, $n = n_0 + n_1$ will be the total of individuals observed and $m_0 \leq n$ will represent the total observations without ties. They first rank the genes accordingly to ABCR (5).

$$ABCR = \sum_{k=1}^{m_0} |AUC_k - A_k|, \tag{5}$$

where AUC_k is the partial area under a ROC curve between the consecutive abscissa points for $k = 1, \ldots, m_0$ computed according to the standard trapezoidal rule, and $A_k = \frac{2k-1}{2m_0^2}$ represent the partial area of the chance line. The first g genes correspond to a False Discovery Rate defined by the user. TNRC statistic is used to test for not proper ROC curves:

$$TNRC = \sum_{k=1}^{m_0} |AUC_k - A_k| - |AUC - 0.5| \tag{6}$$

where AUC is the area under the empirical ROC curve. Not proper ROC curves are identified by high values of the TNRC statistic.

Author's contributions

CSF, a PhD student developed and implemented the method under the guidance of her advisors MAAT and LS. All authors discussed and approved the paper.

Acknowledgements

Algorithm implementation

R source code for the implementation of this algorithm is available upon request.

References

1. Horn T, Sandmann T, Fischer B, Axelsson E, Huber W, Boutros M: **Mapping of signalling networks through synthetic genetic interaction analysis by RNAi**. *Nature Methods* 2011, **8**(4):341–349.

2. Xu Z, Wei W, Gagneur J, Clauder-Munster S, Smolik M, Huber W, Steinmetz L: **Antisense expression increases gene expression variability and locus interdependency**. *Molecular Systems of Biology* 2011, **7**:1–10.

3. Mancera E, Bourgon R, Huber W, Steinmetz LM: **Genome-wide survey of post-meiotic segregation during yeast recombination**. *Genome Biol* 2011, **12**:R36.

4. Thomsen S, Anders S, Janga SC, Huber W, Alonso CA: **Genome-wide analysis of mRNA decay patterns during early Drosophila development**. *Genome Biol* 2010, **11**:R93.

5. Parodi S, Pistoia V, Muselli M: **Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments**. *BMC Bioinformatics* 2008, **9**:410.

6. Alizadeh AA, Elsen MB, Davis E, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Marti GE, Moore T, Hudson Jr J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* 2000, **403**:503–511.

7. Li L, Chaudhuri A, Chant J, Tang Z: **PADGE: analysis of heterogeneous patterns of differential gene expression**. *Physiol Genomics* 2007, **32**:154–159.

8. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *J Am Stat Assoc* 2002, **97**(457).

9. Jeffery IA, Higgins DG, Culhane AC: **Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data**. *BMC Bioinformatics* 2006, **7**:359.

10. Pepe MS, Longton G, Anderson GL, Schummer M: **Selecting differentially expressed genes from microarray experiments**. *Biometrics* 2003, **59**:133–142.

11. Parodi S, Muselli M, Fontana V, Bonassi S: **ROC curves are a suitable and flexible tool for the analysis of gene expression profiles**. *Cytogenet Genome Res* 2003, **101**:90–91.

12. Pepe MS: *The statistical evaluation of medical tests for classification and prediction.* Oxford (UK): Oxford University Press 2003.

13. Metz CE, Pan X: **Proper binormal ROC curves: Theory and maximum-likelihood estimation**. *J Math Psychol* 1999, **43**:1–33.

14. Dorfman DD, Berbaum KS, Brandser EA: **A contaminated binormal model for ROC data: Part I. Some interesting examples of binormal degeneracy**. *Acad Radiol* 2000, **7**(6):420–426.

15. Bradley EL: **Overlapping Coefficient**. In *Encyclopedia of Statistical Sciences, Volume 6.* Edited by Ktoz S, Johnson NL, New York: Chapman and Hall 1985:546–547.

16. Inman HF, Bradley EL: **The overlapping coefficient as a measure of agreement between two probability distributions and point estimation of the overlap of two normal densities**. *Commun Statist Theory Methods* 1989, **18**(10):3851–3872.

17. Hanley JA, McNeill BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve**. *Radiology* 1982, **143**:29–36.

18. Rosenblatt M: **Remarks on some nonparametric estimates of a density function**. *Ann Math Statist* 1956, **27**(3):832–837.

19. **Open-source R software** [http://www.r-project.org/].

20. **Bioconductor: Open-source software for Bioinformatics** [http://www.bioconductor.rg/].

21. Breitling R, Herzyk P: **Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data**. *J Bioinform Comput Biol* 2005, **3**(5):1171–1189.

22. Affymetrix: *Gene chip analysis suite user guide.* Santa Clara: Affymetrix 1999.

23. Kadota K, Nakai Y, Shimizu K: **A weighted average difference method for detecting differencially expressed genes from microarray data**. *Algorithm Mol Biol* 3:8.

24. Smyth GK: **Limma linear models for microarray data**. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, New York: Springer 2005:397–420.

25. Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M: **Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments**. *BMC Bioinformatics* 2006, **7**:538.

26. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci USA* 2001, **98**(9):5116–5121.

27. Silverman BW: *Density estimation for statistics and data analysis*. New York: Chapman and Hall 1986.

28. Venables WN, Ripley BD: *Modern applied statistics with S*. Springer 2002.

29. Bamber D: **The area above the ordinal dominance graph and the area below the receiver operating graph**. *J Math Psychol* 1975, **12**:387–415.

Figures
Figure 1 - Relationship between densities and ROC curves considering equal variances on both groups

Probability density functions of gene expression values of two groups and their corresponding empirical ROC curves, where Y is a random variable which represents the expression values under the experimental condition and X a random variable which represents the expression values for the control group. It was considered the same classification rule in all ROC plots, i.e., high values of the decision variable correspond to positive regulation. Density plots were obtained using kernel density estimation from two samples of size 100 simulated from normal distributions. A)$X \sim N(20, 4), Y \sim N(30, 4)$ B)$X \sim N(20, 4), Y \sim N(22, 4)$ C)$X \sim 0.5N(20, 2) + 0.5N(20, 2), Y \sim N(0, 11)$ D)$X \sim N(0, 11), Y \sim 0.5N(20, 2) + 0.5N(20, 2)$ E)$X \sim N(30, 4), Y \sim N(20, 4)$.

Figure 2 - Relationship between densities and ROC curves, considering different variances and similar means on both groups

Probability density functions of gene expression values of two groups and their corresponding empirical ROC curves, where Y is a random variable which represents the expression values under the experimental group and X is a random variables which represents the expression values for the control group. It was considered the same classification rule in all ROC plots, i.e., high values of the decision variable correspond to positive regulation. Density plots were obtained using kernel density estimation from two samples of size 100 simulated from normal distributions. A)$X \sim N(20, 15), Y \sim N(20, 60)$ B)$X \sim N(20, 40), Y \sim N(20, 5)$.

Figure 3 - Pseudo code for non-parametric estimation of OVL

Figure 4 - Pseudo code to select differentially expressed genes based on AUC and OVL

Figure 5 - Arrow Plot of Lymphoma data

It was considered an AUC$\geq$ 0.9 and OVL$<$ 0.5 to select up-regulated genes, corresponding to red dots on the plot. To select down-regulated genes an AUC$\leq$ 0.1 and OVL$<$ 0.5 was considered, corresponding to blue dots on the plot. To select special genes it was considered an OVL$<$ 0.5 and $0.4 < AUC < 0.6$. Orange dots corresponds to a bimodal or multimodal density in the experimental group, cyan dots correspond to a bimodal or multimodal density in the control group and green dots correspond to a bimodal or multimodal density in both groups.

Figure 6 - Kernel density plots and empirical ROC plots

Kernel density estimate of the 20 special selected genes expression values, where red densities represent the experimental sample and black densities represent the control sample. The x-axis is on log base 2 scale. From left to the right, each plot pair correspond to densities and respective empirical ROC curve of the gene ID's: GENE1141X GENE3521X, GENE3547X, GENE3473X, GENE2547X, GENE2519X, GENE1877X, GENE3343X, GENE3322X, GENE3323X, GENE3389X, GENE3388X, GENE3909X, GENE2887X, GENE2778X, GENE463X, GENE1004X, GENE3407X, GENE75X, GENE1817X.

Figure 7 - Arrow plot of simulated data

Orange dots correspond to truly no differentially expressed genes, red dots to truly up-regulated genes, blue dots to truly down-regulated genes and green dots to truly special genes. It was considered an AUC$\geq$ 0.9 and an OVL$<$ 0.5 to select up-regulated genes. To select down-regulated genes an AUC$\leq$ 0.1 and an OVL$<$ 0.5 were considered and to select differentially expressed genes with bimodal or multimodal densities it were considered an OVL$<$ 0.5 and $0.4 < AUC < 0.6$.

Figure 8 - Comparison of estimated and exact OVL values.

The exact values of OVL (red line) were obtained from two normal distributions with means 20 and 10 and the same variances equal to 16. The estimated OVL values were obtained from two kernel density estimates generated from two samples of size 100, simulated from normal distributions with means 20 and 10 and variances equal to 16 (dots).

Tables

Table 1 - **List with the notation used in Algorithm 1. Symbols are listed in order of appearance.**

Symbol	Definition
G^A, G^B	Kernel density coordinates of samples A and B
A, B	pairs of coordinates of samples A and B that will be used to estimate OVL
$G_x^A[i], G_y^A[i]$	index a pair of coordinates of G^A, where $G_x^A[i]$ is the abscissa and $G_y^A[i]$ is the correspondent ordinate (same to G^B and G)
$\sharp(.)$	total number of pairs of coordinates
G	ordered list of points resulting from the union of A and B
P	union of G with new pairs of coordinates, which correspond to jump points between densities
$G[i]$	indexes a pair of coordinates of G
x_{new}	new abscissa
y_{new}	new ordinate
F	final list of pairs of coordinates to estimate OVL
OVL	overlapping coefficient between two kernel densities

Table 2 - **List of functions used in Algorithm 1. Functions are listed in order of appearance**

Function	Definition
xMatch(abscissa,list)	if there is more than one equal abscissa on the list, returns the pair of coordinates corresponding to the one which has the minimum ordinate
ordinate(abscissa,ordinate)	returns the ordinate of a pair of coordinates
xPrev(abscissa,list)	returns the pair of coordinates immediately preceding the abscissa in the list
xNetx(abscissa,list)	returns the pair of coordinates immediately after the abscissa in the list
Union(list,list)	joins lists
order(list)	orders a list in increasing order of abscissas
abscissa(abcissa,ordinate)	returns the abscissas of a pair of coordinates
trapez(list)	trapezoidal rule for area estimation

Table 3 - **List with the notation used in Algorithm 2. Symbols are listed in order of appearance.**

Symbol	Definition
X	$p \times n$ matrix, corresponding to sample A with columns representing arrays and rows representing genes
Y	$p \times m$ matrix, corresponding to sample B with columns representing arrays and rows representing genes
UP	up-regulated genes list
DOWN	down-regulated genes list
$X[i]$, $Y[i]$	indexes a gene (row of the matrix)
$k1$, $k2$, $k3$ $k4$, w	arbitrary thresholds
S^A, S^B	kernel density coordinates of subsamples of genes from samples A and B
$S^{A[j]}$	indexes a gene of the subsample S from sample A
$S_y^{A[j]}[i]$	indexes a ordinate of a gene j of the subsample S from sample A
$Bim_X[j]$	indexes a gene with bimodal or multimodal kernel density from sample A
$Bim[j]$	indexes a gene with bimodal or multimodal kernel density
SPECIAL	special genes list

Table 4 - List of functions used in Algorithm 2. Functions are listed in order of appearance

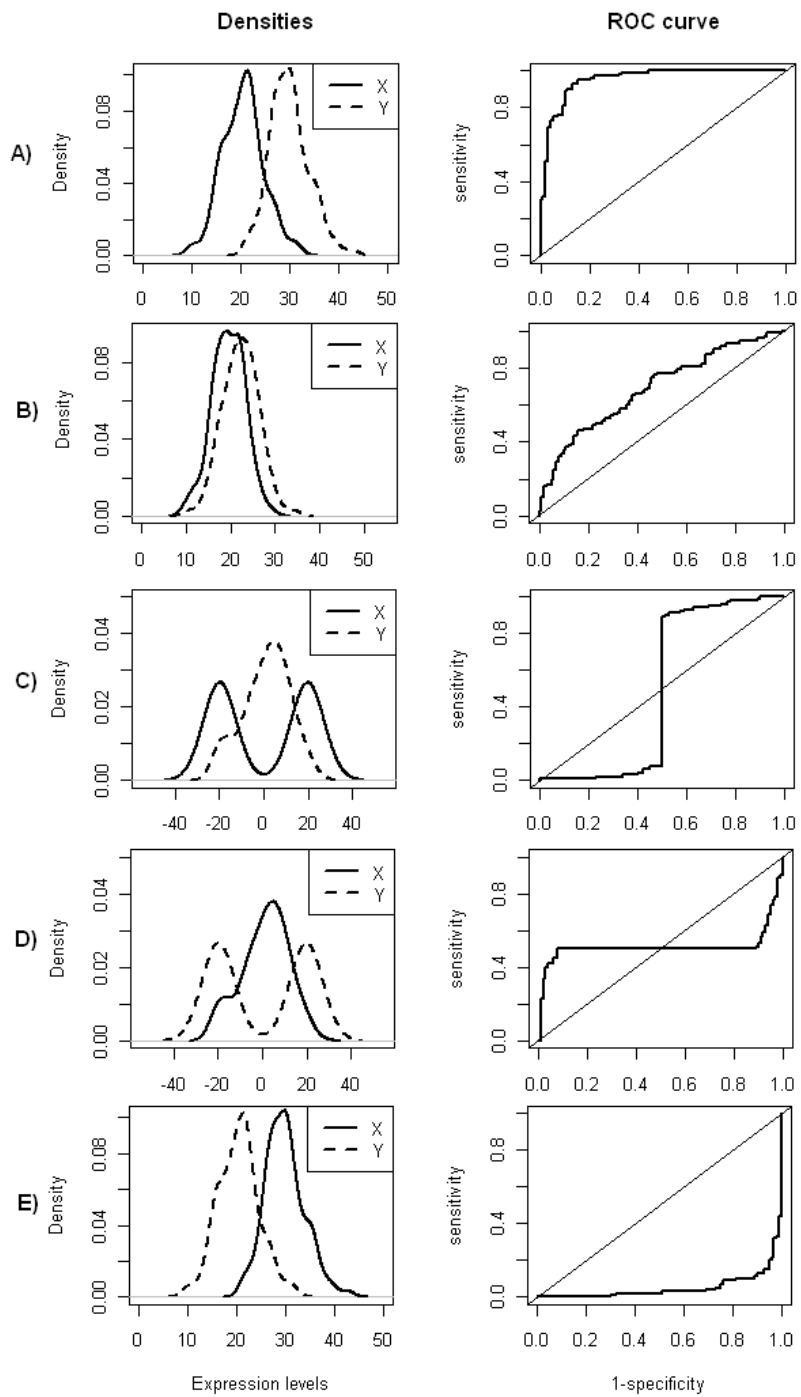| Function	Definition
`AUC(list,list)`	Area above the ROC curve estimated by the trapezoidal rule
`OVL(list,list)`	overlapping coefficient estimated by Algorithm 1
`kernel(list)`	kernel density estimation
`rank(list)`	returns the ranks of a list

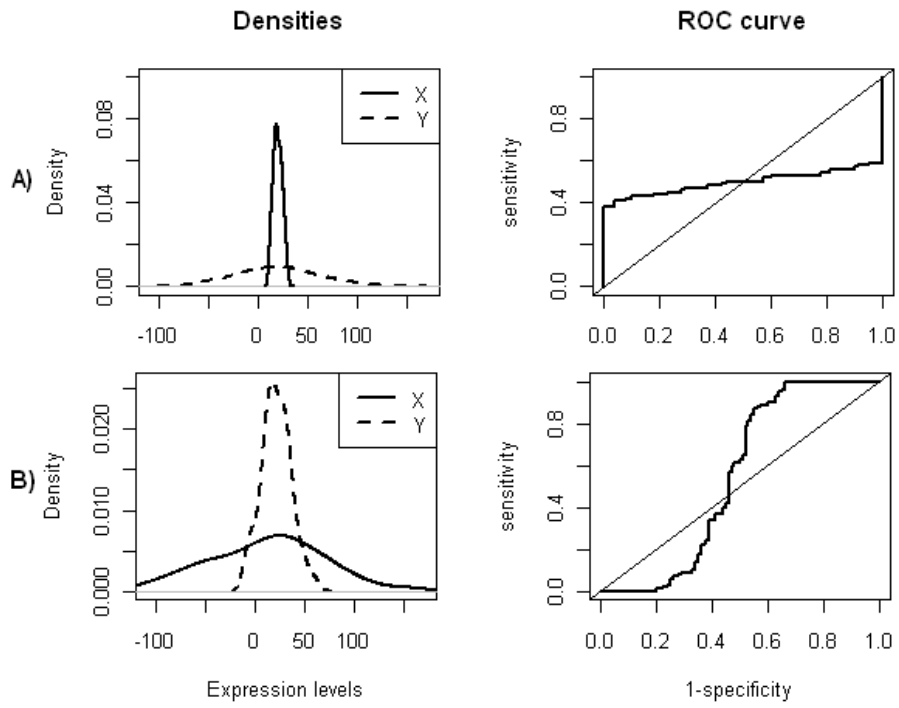**Table 5 - AUC and OVL values and bimodality group identification of the 20 selected special genes**

Special genes were selected using OVL< 0.5 and 0.4<AUC<0.6. E: bimodality in experimental group, C: bimodality in control group and B: bimodality in both groups. Genes are ordered by ascending order of OVL.

Gene ID	Gene name	OVL	AUC	Group
GENE3323X	*BCL7A*	0.389	0.477	B
GENE3473X	*Unknown*	0.399	0.407	B
GENE1877X	*Unknown*	0.428	0.421	B
GENE3388X	*Immunoglobulin J chain*	0.432	0.529	B
GENE1141X	*MAPKKK5*	0.443	0.571	E
GENE3521X	*Similar to KIAA0050*	0.446	0.593	B
GENE3407X	*Histone deacetylase 3*	0.453	0.543	B
GENE75X	*VRK2 kinase*	0.457	0.546	C
GENE2519X	*Unknown*	0.461	0.529	E
GENE3343X	*LR11*	0.461	0.543	B
GENE1817X	*BL34*	0.472	0.586	B
GENE3389X	*Immunoglobulin J chain*	0.476	0.475	E
GENE3909X	*Placental bikunin*	0.492	0.463	C
GENE2887X	*LBR*	0.492	0.486	E
GENE3547X	*Immunoglobulin kappa light chain*	0.493	0.413	B
GENE1004X	*BNIP3*	0.494	0.511	B
GENE2547X	*CLK3 kinase*	0.495	0.500	B
GENE2778X	*DNA Ligase III*	0.496	0.536	B
GENE3322X	*BCL7A*	0.498	0.532	E
GENE463X	*PARP*	0.499	0.461	B

Table 6 - Biological description of the 20 selected special genes

Gene name	Description
BCL7A	Chromosome 12q24.1 is the site of a recurrent breakpoint in high-grade B-cell non-Hodgkin lymphoma of the amino-terminus of BCL7A defined a new mechanism in the pathogenesis of a subset of high-grade B-cell non-Hodgkin lymphoma.
Immunoglobulin J chain	The J chain is found in pentameric IgM, and it is also important as a secretory immunoglobulin.
MAPKKK5	JUN N-terminal kinases (JNKs) are MAPKs that stimulate transcriptional activity of JUN in response to growth factors, proinflammatory cytokines, and certain environmental stresses.
Similar to KIAA0050	ACAP1 and ACAP2 were recruited to platelet-derived growth factor (PDGF)-induced dorsal membrane ruffles in NIH 3T3 mouse fibroblasts, and overexpression inhibited ruffle formation.
Histone deacetylase 3	Acetylation and deacetylation of histone proteins alters chromosome structure and affects transcription factor access to DNA.
VRK2 kinase	Genes involved in the regulation of cell division.
LR11	Is genetically associated with late-onset Alzheimer's disease and is thought to be involved in neurodegenerative processes.
BL34	Sequence of a human B-cell specific immediate early gene, that is inducible in response to several B-cell activation signals.
Placental bikunin	Is a potent inhibitor of serine proteases involved in blood coagulation and fibrinolysis such as human plasmin, plasma and tissue kallikrein, and factor XIa.
LBR	Encodes the lamin B receptor, an inner nuclear membrane protein that binds lamin B. Since the lamin B receptor may be a sterol reductase, loss of most LBR expression might lead to changes in sterol metabolism that cause developmental abnormalities.
Immunoglobulin kappa light chain	Small polypeptide subunit of an antibody. Antibodies are produced by B lymphocytes, each expressing only one class of light chain (Kappa or lambda).
BNIP3	Activates expression of the gene encoding NIP3, in turn primes cells for apoptosis under conditions of persistent oxygen deprivation.This pathway may play play a role in cell death resulting from cerebral and myocardial ischemia.
CLK3 kinase	A family of enzymes that catalyze the phosphorylation of proteins at serine and threonine residues.
DNA Ligase III	Seals DNA strand breaks that arise during the process of meiotic recombination in germ cells and as a consequence of DNA damage in somatic cells.
PARP	Is required for cellular repair. Inhibitors of this enzyme potentiate the lethal effects of noxious agents.

Densities

ROC curve

A)

B)

C)

D)

E)

Expression levels

1-specificity

input : G^A, G^B
output: OVL kernel-based estimation

1 $i \leftarrow 1$;
2 $A \leftarrow$ empty;
3 **while** $i <= \sharp(G^A)$ **do**
4 $\quad$ $x_1 \leftarrow G_x^A[i]$;
5 $\quad$ $y_1 \leftarrow G_y^A[i]$;
6 $\quad$ **if** { [xMatch$(x_1, G^B) \neq empty \wedge y_1 \leq$ordinate(xMatch$(x_1, G^B)$)] $\vee$
7 $\quad$ [xMatch$(x_1, G^B) = empty \wedge$ xPrev$(x_1, G^B) \neq empty \wedge$ xNext$(x_1, G^B) \neq empty$
$\quad\quad \wedge y_1 \leq$ordinate(xPrev$(x_1, G^B)$) $\wedge y_1 \leq$ ordinate(xNext(x_1, G^B))] } **then**
8 $\quad\quad$ $A \leftarrow (G_x^A[i], G_y^A[i])$;
9 $\quad$ **end**
10 $\quad$ $i \leftarrow i + 1$;
11 **end**
12 $i \leftarrow 1$;
13 $B \leftarrow$ empty;
14 **while** $i <= \sharp(G^B)$ **do**
15 $\quad$ $x_2 \leftarrow G_x^B[i]$;
16 $\quad$ $y_2 \leftarrow G_y^B[i]$;
17 $\quad$ **if** { [xMatch$(x_2, G^A) \neq empty \wedge y_2 \leq$ordinate(xMatch$(x_2, G^A)$)] $\vee$
18 $\quad$ [xMatch$(x_2, G^A) = empty \wedge$ xPrev$(x_2, G^A) \neq empty \wedge$ xNext$(x_2, G^A) \neq empty$
$\quad\quad \wedge y_2 \leq$ ordinate(xPrev(x_2, G^A)) $\wedge y_2 \leq$ ordinate(xNext(x_2, G_y^A))] } **then**
19 $\quad\quad$ $B \leftarrow (G_x^B[i], G_y^B[i])$;
20 $\quad$ **end**
21 $\quad$ $i \leftarrow i + 1$;
22 **end**
23 $G \leftarrow$ order(Union(A, B));
24 $i \leftarrow 1$;
25 P$\leftarrow$ empty;
26 **while** $i \leq \sharp(G) - 1$ **do**
27 $\quad$ **if** $(G[i] \in A \wedge G[i+1] \in B)$ **then**
28 $\quad\quad$ $x_1 \leftarrow G_x[i]$;
29 $\quad\quad$ $y_1 \leftarrow G_y[i]$;
30 $\quad\quad$ $x_4 \leftarrow G_x[i+1]$;
31 $\quad\quad$ $y_4 \leftarrow G_y[i+1]$;
32 $\quad\quad$ **if** $G[i] \in A$ **then** $x_2 \leftarrow$abscissa(xNext(x_1, G^A));
33 $\quad\quad$ **else** $x_2 \leftarrow$ abscissa(xNext(x_1, G^B));
34 $\quad\quad$ **if** $G[i] \in A$ **then** $y_2 \leftarrow$ordinate(xNext(x_1, G^A));
35 $\quad\quad$ **else** $y_2 \leftarrow$ordinate(xNext(x_1, G^B));
36 $\quad\quad$ **if** $G[i+1] \in A$ **then** $x_3 \leftarrow$abscissa(xPrev(x_4, G^A));
37 $\quad\quad$ **else** $x_3 \leftarrow$ abscissa(xPrev(x_4, G^B));
38 $\quad\quad$ **if** $G[i+1] \in A$ **then** $y_3 \leftarrow$ordinate(xPrev(x_4, G^A));
39 $\quad\quad$ **else** $y_3 \leftarrow$ ordinate(xPrev(x_4, G^B));
40 $\quad\quad$ $D \leftarrow (x_1 - x_2) \times (y_3 - y_4) - (y_1 - y_2) \times (x_3 - x_4)$;
41 $\quad\quad$ **if** $D \neq 0$ **then**
42 $\quad\quad$ $x_{new} \leftarrow (1/D) \times [(x_3 - x_4)(x_{12} - y_1 \times x_2) - (x_1 - x_2)(x_3 \times y_4 - y_3 \times x_4)]$;
43 $\quad\quad$ $y_{new} \leftarrow (1/D) \times [(y_3 - y_4)(x_1 \times y_2 - y_1 \times x_2) - (y_1 - y_2)(x_3 \times y_4 - y_3 \times x_4)]$;
44 $\quad\quad$ **else** P$\leftarrow$ (P, (x_1, y_1), (x_{new}, y_{new}));
45 $\quad$ **else**
46 $\quad\quad$ P$\leftarrow$ (P,$(G_x[i], G_y[i])$);
47 $\quad$ **end**
48 $\quad$ $i \leftarrow i + 1$;
49 **end**
50 F$\leftarrow$ order(Union$(P, (G_x[\sharp(G)], G_y[\sharp(G)]))$);
51 OVL$\leftarrow$Trapez(F)

input : X, Y
output: Genes with differential expression

1 UP$\leftarrow$ empty;
2 DOWN$\leftarrow$ empty;
3 K$\leftarrow$ empty;
4 $i \leftarrow 1$
5 **for** $i \leftarrow p$ **do**
6   **if** `AUC`$(X[i], Y[i]) > k_1 \wedge$ `OVL`(`kernel`$(X[i])$,`kernel`$(Y[i])) < w$ **then**
7 | UP$\leftarrow i$;
8
9   **if** `AUC`$(X[i], Y[i]) < k_2 \wedge$ `OVL`(`kernel`$(X[i])$,`kernel`$(Y[i])) < w$ **then**
10 | DOWN$\leftarrow i$;
11
12   **if** `AUC`$(X[i], Y[i]) < k_3 \wedge$ `AUC`$(X[i], Y[i]) > k_4 \wedge$
       `OVL`(`kernel`$(X[i])$,`kernel`$(Y[i])) < w$ **then**
13 | K$\leftarrow i$;
14
15 **end**
16 $z_1 \leftarrow$ empty;
17 $z_2 \leftarrow$ empty;
18 $Bim_X \leftarrow$ empty;
19 $Bim_Y \leftarrow$ empty;
20 SPECIAL$\leftarrow$ empty;
21 $j \leftarrow 1$ **for** $j \in K$ **do**
22 $S^A \leftarrow$ `order`(`kernel`$(X[j])$);
23 $S^B \leftarrow$ `order`(`kernel`$(Y[j])$);
24   **while** $i \leq \sharp(S^{A[j]})$ **do**
25   **if** $S_y^{A[j]}[i] \leq S_y^{A[j]}[i+1]$ **then**
26 | $z_1^{[j]} \leftarrow 1$;
27   **else**
28 | $z_1^{[j]} \leftarrow 0$;
29   **end**
30 $r_1^{[j]} \leftarrow$ `rank`($z_1^{[j]}[i]$)-`rank`($z_1^{[j]}[i+1]$);
31   **for** $z_1^{[j]} \neq 1$ **do**
32   **if** $\exists' r_1^{[j]}[i] > 1$ **then**
33 | $Bim_X[j] \leftarrow$ TRUE ;
34
35   **end**
36   **end**
37   **while** $i \leq \sharp(S^{B[j]})$ **do**
38   **if** $S_y^{B[j]}[i] \leq S_y^{B[j]}[i+1]$ **then**
39 | $z_2^{[j]} \leftarrow 1$;
40   **else**
41 | $z_2^{[j]} \leftarrow 0$;
42   **end**
43 $r_2^{[j]} \leftarrow$ `rank`($z_2^{[j]}[i]$)-`rank`($z_2^{[j]}[i+1]$);
44   **for** $z_2^{[j]} \neq 1$ **do**
45   **if** $\exists' r_2^{[j]} > 1$ **then**
46 | $Bim_Y[j] \leftarrow$ TRUE ;
47
48   **end**
49   **end**
50   **if** $Bim_X[j] = TRUE \vee Bim_Y[j] = TRUE$ **then**
51 | $Bim[j] =$TRUE ;
52
53   **if** $Bim[j] = TRUE$ **then**
54 | SPECIAL$\leftarrow j$
55
56 **end**