

## A class comparison method with filtering-enhanced variable selection for high-dimensional data sets<sup>‡</sup>

Lara Lusa<sup>1,\*</sup>, Edward L. Korn<sup>2</sup> and Lisa M. McShane<sup>2</sup>

<sup>1</sup>*Department of Medical Informatics, University of Ljubljana, Slovenia*

<sup>2</sup>*Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892-7434, U.S.A.*

### SUMMARY

High-throughput molecular analysis technologies can produce thousands of measurements for each of the assayed samples. A common scientific question is to identify the variables whose distributions differ between some pre-specified classes (i.e. are differentially expressed). The statistical cost of examining thousands of variables is related to the risk of identifying many variables that truly are not differentially expressed, and many different multiple testing strategies have been used for the analysis of high-dimensional data sets to control the number of these false positives. An approach that is often used in practice to reduce the multiple comparisons problem is to lessen the number of comparisons being performed by filtering out variables that are considered non-informative 'before' the analysis. However, deciding which and how many variables should be filtered out can be highly arbitrary, and different filtering strategies can result in different variables being identified as differentially expressed. We propose the filtering-enhanced variable selection (*FEVS*) method, a new multiple testing strategy for identifying differentially expressed variables. This method identifies differentially expressed variables by combining the results obtained using a variety of filtering methods, instead of using a pre-specified filtering method or trying to identify an optimal filtering of the variables prior to class comparison analysis. We prove that the FEVS method probabilistically controls the number of false discoveries, and we show with a set of simulations and an example from the literature that FEVS can be useful for gaining sensitivity for the detection of truly differentially expressed variables. Published in 2008 by John Wiley & Sons, Ltd.

**KEY WORDS:** multiple testing methods; multivariate permutation methods; high-dimensional data; microarrays; variable filtering

\*Correspondence to: Lara Lusa, Department of Medical Informatics, University of Ljubljana, Slovenia.

†E-mail: Lara.Lusa@mf.uni-lj.si

‡This article is a U.S. Government work and is in the public domain in the U.S.A.

Contract/grant sponsor: Istituto Superiore di Sanità on Oncological Pharmacogenomics—Seroproteomics  
Contract/grant sponsor: AIRC (Associazione Italiana per la Ricerca sul Cancro)

## 1. INTRODUCTION

High-throughput molecular analysis technologies can produce thousands of measurements for each of the assayed samples. For example, gene-expression microarray experiments measure simultaneously the expression of thousands of genes (*variables*), which comprise a 'profile' for each specimen. A common scientific question is whether and how the profiles differ on average between two or more different classes of specimens. Although this question can be asked globally (are the average profiles different between classes?), typically there will be interest in identifying specific variables whose distributions differ between the classes (referred to here as differentially expressed variables); see [1, 2].

One approach to this problem is to perform an appropriate (univariate) statistical test for each variable (e.g. a *t*-test for a two-class comparison), and then, because thousands of variables have been examined, adjust the results to control probabilistically the number or proportion of variables erroneously identified as differentially expressed (false positives or false discoveries). Many different methods have been used with high-dimensional data (see [3] for a review).

In practice, many investigators reduce the multiple testing problem by removing from their data sets those variables that show very little variation in expression across all the specimens regardless of class. For instance, Yamanaka *et al.* [4] used a univariate proportional hazards model to identify the genes associated with survival in a cohort of malignant glioma patients; only the genes whose expression differed by at least 1.5-fold from the median in at least 20 per cent of the arrays were retained in their analysis. Fält *et al.* [5] identified the genes that were differentially expressed between the irradiated and the non-irradiated human lymphocytes excluding the genes with low variation across their data set (less than 3-fold ratio between maximum and minimum and less than 100 difference between maximum and minimum observed intensities). Additional examples can be found in [6–11]. The rationale behind this practice is that the variables that are filtered out are unlikely to be differentially expressed between classes and, at the same time, their removal might improve the sensitivity of the multiple-comparisons-adjusted analysis. This is because the variability (regardless of class) of the differentially expressed variables is inflated by the differences between classes and therefore, even if these variables have the same intraclass variability as the null variables, they are more likely to be retained in the filtered data sets.

Variable filtering requires two choices: (i) which statistic will be used to rank the variables for filtering purposes (the filtering ranking-statistic; e.g. the sample variance of the variable, the ratio of the 95th to 5th percentile, a max/min-based ratio, or the interquartile range) and (ii) how many variables will be filtered out (the stringency of the filtering; e.g. a pre-specified proportion of the variables should be filtered out, or variables to drop depend on a threshold value for the filtering ranking-statistics, which is chosen *a priori*). These choices are clearly arbitrary and can have a substantial impact on the final results of the analyses, i.e. on the list of variables identified as differentially expressed. For example, the use of a particularly stringent filtering method that filters out many variables, might exclude from consideration some of the truly differentially expressed variables. On the other hand, the use of a less stringent filtering method (which filters out few or none of the variables) might fail to identify some of the truly differentially expressed variables due to the multiple comparisons correction, especially in experiments that do not have high power.

This paper proposes a new multiple testing strategy for identifying differentially expressed variables in high-dimensional data sets, the filtering-enhanced variable selection (FEVS) method. This method identifies differentially expressed variables by combining the results obtained using a variety of filtering methods. The FEVS method is based on the multivariate permutation procedure

to control the number of false discoveries [12] and, therefore, it can be used in all the situations in which multivariate permutation methods are applicable provided that it is feasible to specify a criterion to filter out potentially non-informative variables. This non-parametric procedure accounts for the correlation between the variables and we prove that it controls probabilistically the number of false discoveries.

In Section 2 we describe FEVS and show how to apply it with the class-independent filtering methods. In Section 3 we present a set of limited simulations to show that in a variety of situations FEVS can be useful for gaining sensitivity for the detection of the truly differentially expressed variables, while controlling the number of false discoveries. We apply the method to published data from a breast cancer gene expression microarray study in Section 3.2. We end with a discussion in Section 4, including the possibility of using FEVS with class-dependent filtering.

## 2. METHODS

We restrict attention to the two-class comparison in this section; see the discussion for description of the straightforward extension to other situations. The FEVS method to identify which of the  $k$  variables has class differences extends the multivariate permutation testing (MPT) method described by Korn *et al.* [12]. It is assumed that there exists a  $k_0$ -dimensional subvector of the  $k$ -dimensional vector of variables, which is independent and identically distributed regardless of the class from which the sample came. We denote variables in the subvector as ‘null hypothesis variables’ and note that they each satisfy the standard univariate null hypothesis of no class difference. Without filtering, a  $k$ -dimensional vector  $(p_1, p_2, \dots, p_k)$  of  $p$ -values is calculated for which the  $i$ th component is equal to the univariate  $p$ -value for the  $i$ th variable using a two-sample test (e.g. a  $t$ -test). (Instead of  $p$ -values, any univariate statistic could be used, which ranked the variables according to their ability to discriminate the classes.) The data vectors are then permuted between the two classes and the vector of  $p$ -values is recomputed,  $(p_1^*, p_2^*, \dots, p_k^*)$ . The MPT-based procedure that identifies variables whose original  $p$ -value  $(p_i)$  is less than the  $\alpha$  quantile of the permutation distribution of the  $(u+1)$ st smallest  $p$ -value of  $(p_1^*, p_2^*, \dots, p_k^*)$  will identify more than  $u$  null-hypothesis variables with probability  $\leq \alpha$  [12].

To extend the MPT procedure to incorporate filtering, consider  $S$  different class-independent strategies, each of which filters out variables according to a different rule. Let  $k_s$  be the number of variables that is not filtered out by the  $s$ th filtering criterion. For filtering method  $s$  and variable  $i$  not filtered out by this method, we define a ‘Bonferroni-like adjusted  $p$ -value ranking-statistic’ (BLAPS) to be the product of the observed  $p$ -value on the complete data set  $(p_i)$  and the number of variables not filtered out by that filtering method  $(k_s)$ . Note that the larger the number of variables filtered out by the filtering method, the smaller the required correction applied to the  $p$ -value from the complete data set. When variable  $i$  is filtered out by a given filtering method  $s$ , its BLAPS is set equal to  $k$ . BLAPS are not exactly equivalent to the well-known Bonferroni-adjusted  $p$ -values, since we do not constrain them to take values  $\leq 1$ . We define the quantity  $m_i$  as the minimum BLAPS obtained for the  $i$ th variable when applying the  $S$  filtering methods. The quantity  $m_i$  will be equal to the BLAPS derived from the filtering method that filters out the largest number of variables among those methods that do not filter out the variable  $i$ . The rationale behind the use of the  $m_i$  quantities is to reduce the multiple comparisons problem in a variable-dependent way: variables that are filtered out early (before most of the others) with a less stringent filtering method benefit little from the reduction of the multiple comparisons correction produced by the use of

a filtering method, and their  $p$ -values are adjusted with a large  $k_s$ , while the opposite happens for the variables that are filtered out late (after most of the others) with a more stringent filtering method. However, none of the variables is excluded *a priori* from being identified as differentially expressed by the use of a specific filtering method.

The following generalization of the results given in [12], shows that we can use a class of statistics (that includes the  $m_i$ ) to control probabilistically the number of false positives. The results given in [12] do not directly apply since the statistic  $m_i$  is not just a function of the data for variable  $i$ .

*Proposition*

Let  $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$  and  $\{\mathbf{Y}_1, \dots, \mathbf{Y}_m\}$  be the  $k$ -dimensional data vectors for class 1 and 2, respectively. We assume that a  $k_0$ -dimensional subvector of the  $k$ -dimensional vectors is independent and identically distributed (the ‘null hypothesis variables’). Let

$$W_i = g(\{X_{1i}, \dots, X_{ni}\}, \{Y_{1i}, \dots, Y_{mi}\}, \{\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_m\}) \tag{1}$$

be a function of the set of class 1 data values for variable  $i$ , the set of class 2 data values for variable  $i$ , and the set of data vectors regardless of class, with smaller values of  $W_i$  suggesting a class difference for variable  $i$ . (Expression (1) requires that  $W_i$  does not depend on class labels except for the association of the values of variable  $i$  with class label.) Consider constructing permuted data sets by permuting the class labels, and for each permuted data set, calculating the  $W$ ’s on all the variables, say,  $W_1^*, W_2^*, \dots, W_k^*$ . Let  $W_{(1)}^*, W_{(2)}^*, \dots, W_{(k)}^*$  denote the ordered  $W^*$ ’s, and let

$$c_{\alpha,u} = \text{MAX}(c | P(W_{(u)}^* < c | \{x_1, \dots, x_n, y_1, \dots, y_m\}) \leq \alpha)$$

where  $P(\bullet | \{x_1, \dots, x_n, y_1, \dots, y_m\})$  refers to the probability under the permutation distribution. (The quantity  $c_{\alpha,u}$  is, using one standard definition of quantiles for discrete distributions [13], the  $\alpha$  quantile of the distribution of  $W_{(u)}^*$  conditional on  $\{x_1, \dots, x_n, y_1, \dots, y_m\}$ .) The procedure that identifies all variables with  $W_i < c_{\alpha,u}$  will identify more than  $u$  null-hypothesis variables with probability  $\leq \alpha$ . In other words, the procedure will control the number of false-positive results (false discoveries) with  $1 - \alpha$  confidence.

*Proof*

Let  $I \subseteq \{1, \dots, k\}$  be the set of indices corresponding to the null-hypotheses variables, let  $W_{(1)}^0 \leq W_{(2)}^0 \leq \dots \leq W_{(k_0)}^0$  be the ordered  $W$  statistics on the original (unpermuted) data set restricted to  $i \in I$ , let  $W_{(1)}^{0*} \leq W_{(2)}^{0*} \leq \dots \leq W_{(k_0)}^{0*}$  be the ordered  $W$  statistics on a permuted data set restricted to  $i \in I$ , and let

$$c_{\alpha,u}^0 = \text{MAX}(c | P(W_{(u)}^{0*} < c | \{x_1, \dots, x_n, y_1, \dots, y_m\}) \leq \alpha)$$

Note that though  $c_{\alpha,u}^0$  is unknown to us, we do know that  $c_{\alpha,u}^0 \geq c_{\alpha,u}$  because  $W_{(u)}^{0*} \geq W_{(u)}^*$  (the  $\{W^{0*}$ ’s} being a subset of the  $\{W^*$ ’s}). The proof that the probability that  $u$  or more null hypotheses variables are identified is  $< \alpha$  is given by

$$\begin{aligned} &P(u \text{ or more null hypotheses variables identified}) \\ &= P(W_{(u)}^0 < c_{\alpha,u}) \\ &= E[P(W_{(u)}^0 < c_{\alpha,u} | \{\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_m\})] \end{aligned}$$

$$\begin{aligned}
&\leq E[P(W_{(u)}^0 < c_{\alpha,u}^0 | \{\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_m\})] \\
&= E[P(W_{(u)}^{0*} < c_{\alpha,u}^0 | \{\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_m\})] \\
&\leq E(\alpha) = \alpha
\end{aligned}$$

where the penultimate equality follows because the constraint specified by (1) on the  $W$  functions ensures that

$$P(W_{(u)}^0 < c | \{\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_m\}) = P(W_{(u)}^{0*} < c | \{\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{Y}_1, \dots, \mathbf{Y}_m\}) \quad (2)$$

In our application, the  $W_i$  are the  $m_i$ , which satisfy (1). In practice, we use the following algorithm for controlling the number of false discoveries to be less than or equal to a specified number  $u$  with  $1 - \alpha$  confidence. In this algorithm, the univariate  $p$ -values of the original and of the permuted data sets in the Korn *et al.* [12] algorithm are replaced by the  $m_i$  quantities.

### 2.1. FEVS (to control for $\leq u$ false positives)

- (0) Calculate the  $m_i$  quantities for the original data set.
- (1) Initialize the counters  $\text{COUNT}_i = 0$  for  $i = 1, \dots, k$ .
- (2) Choose a random permutation of the sample profiles consistent with the experimental design. Denote the univariate  $p$ -values for the variables from this permutation by  $(p_1^*, p_2^*, \dots, p_k^*)$ . Apply the  $S$  filtering methods to each permuted data set and obtain for each of the variables the  $m_i^*$  quantities, defined as described for the original data.
- (3) Let  $q^* = [\{m_1^*, m_2^*, \dots, m_k^*\}]_{(u+1)}$ , where the notation  $[A]_{(j)}$  is defined as the  $j$ th smallest of the elements of the set  $A$ .
- (4) If  $m_i \geq q^*$  then  $\text{COUNT}_i = \text{COUNT}_i + 1$  for  $i = 1, \dots, k$ .
- (5) Repeat steps 2–4  $B$  times.
- (6) Define the ‘adjusted  $p$ -values’  $\hat{p}_i = (1 + \text{COUNT}_i) / (1 + B)$  for  $i = 1, \dots, k$ .
- (7) If  $u > 0$ , let  $J$  be the set of indices of the  $u$  smallest  $m_i$  values,  $i = 1, \dots, k$ , and let  $\hat{p}_i = 0$  for  $i \in J$ .

The FEVS method calls the differentially expressed variables for which  $\hat{p}_i \leq \alpha$ . If the sample sizes are small enough so that all the permutations can be enumerated, then all of the permutations except the one corresponding to the observed data should be used in Steps 2–4 and  $B$  equals the total number of permutations minus one.

Our procedure is not tailored to a specific filtering method and in principle any  $S$  class-independent filtering method can be used. As a particular case of FEVS, the set of the  $S$  considered filtering methods can be chosen to include those based on the same filtering ranking-statistic that filter out variables with all the possible degrees of stringency, i.e. filtering out, respectively, none of the variables, the variable with the least variation, the two variables with the least variation,  $\dots$  and the  $k - 1$  variables with the least variation (retaining only the variable with most variation). This particular choice of the set of  $S$ , consisting of  $k$  nested filtering methods, is particularly appealing because it avoids the need to specify how many filtering methods should be considered and which stringency should be used. At the same time, it turns out to be particularly convenient from a computational point of view. The  $m_i$  can be calculated straightforwardly, multiplying the original  $p$ -values by the rank of the variables according to the filtering ranking-statistic (the higher the ranking of the variable, the more stringent filtering method is needed to filter out the

variable). In general, we can define  $m_i = p_i \cdot \text{rank}_i$ , where  $\text{rank}_i$  is the rank of the  $i$ th variable according to the filtering ranking-statistic (with smaller ranks meaning larger variability). This simplifies the computations, because only the original  $p$ -values and the ranking of the variable according to the selected filtering ranking-statistic are needed to derive the  $m_i$ , and it is no longer necessary to derive explicitly which variables are filtered out by each of the  $S$  filtering methods. The same simplification applies for deriving the  $m_i^*$  within each permuted data set: the ranking of the variables according to the filtering ranking-statistic does not change between the original and the permuted data sets because we use class-independent filtering ranking-statistics; therefore, for each permutation one needs to derive only the  $p$ -values for all the variables.

We conducted several simulation studies to assess the performance of this version of FEVS, where the interquartile range was used to rank the variables for filtering. In a set of preliminary simulations we observed a loss of sensitivity when all the possible stringencies of the filtering methods were considered, compared with the situations in which the filtering methods that filtered out all but few variables were excluded. In practice, when 10 000 or 50 000 variables were considered, excluding from the set of filtering methods those that retained 100 or less variables proved to be effective in avoiding this loss of sensitivity (data not shown); therefore, this version of FEVS was used in all the simulations that are presented. Once again, the adaptation of the original FEVS method to take into account this modification of the algorithm is straightforward, with  $m_i = p_i \cdot \text{rank}_i$  if  $\text{rank}_i > 100$  and  $m_i = p_i \cdot 100$ , otherwise.

An R package for performing the FEVS method is available at <http://linus.nci.nih.gov/Data/LusaL/bioinfo/> and at CRAN (<http://www.r-project.org/>).

### 3. RESULTS

#### 3.1. Simulation results

We conducted several simulation studies to assess the performance of FEVS. Unless otherwise noted, all the simulations that are presented used the version of FEVS in which the variables were ranked based on their interquartile range and all but the filtering methods that retained less than 100 variables were considered. We compared the class comparison results from FEVS with those obtained (i) without filtering out any of the variables from the data sets, (ii) with filtering out a fixed percentage of the variables (per cent) based on the same filtering ranking-statistic used with FEVS and (iii) with a naïve filtering method ('longest list filtering') that selects a fixed-percentage filtering method after considering many of them, by choosing the one that identifies the longest list of differentially expressed variables. All the analyses with filtering methods different from FEVS are based on the MPT-based procedure [12], in which we control for the same number of false positives and with the same confidence as we do for FEVS.

Even though FEVS can be used in more complex settings, we considered in all the simulations and examples a simple two-class comparison problem, in which univariate (unadjusted)  $p$ -values comparing the variable expression between the two groups were based on parametric two-sample  $t$ -tests with pooled variances. Each simulation was repeated 10 000 times. For purposes of computational feasibility, 99 permutations were performed at each iteration. We controlled the number of errors to be less than or equal to  $u=0$  or  $u=10$ , with 95 per cent confidence.

In the simulation studies, the expression of 50 000 variables for two groups of 5, 20 or 50 samples each was simulated independently from a multivariate Gaussian distribution. Variances

for each variable were drawn from an inverse gamma distribution with shape parameter  $a=3$  and scale parameter  $b=1$ , as in [14].

*3.1.1. Global null.* In the first set of simulations (Table I), zero means were used for all of the variables in both the classes (global null case) and the variables were uncorrelated. Table I reports the proportion of the simulations for which the number of false discoveries was bigger than  $u$  ( $u=0$  and  $u=10$ ) for FEVS and for the other filtering methods considered (nominal proportion=0.05). The number of samples in each group was 5, 20 and 50. The FEVS method satisfied the targeted 95 per cent confidence and so did, as expected, the class-independent fixed-percentage filtering methods considered. The naïve ‘longest-list’ procedure identified a number of (false positive) variables higher than the number of allowed errors in more than 5 per cent of the cases. For example, when considering 5 samples per group and allowing for 10 errors, more than 10 variables were identified in 22 per cent of the simulations.

*3.1.2. Alternative case.* In the second set of simulations (Table II), 300 out of the 50 000 variables were differentially expressed between the two classes. The difference in the means between the two classes for these 300 variables ranged between 0.6 and 3.5 (10 variables in increments of 0.1 units between 0.6 and 3.5). For the case of five samples per group, the first two columns of the first panel in Table II show the increased sensitivity of FEVS in detecting differentially expressed variables compared with no filtering when the variables were uncorrelated. In this simulation, fixed-50 per cent filtering was similar in performance to no filtering and the FEVS performance was similar to fixed-97.5 per cent filtering, identifying more differentially expressed variables among those with higher mean shifts (Table II and Figure 1(a)). In general, we observed that a larger number of variables were identified using more stringent filtering criteria. However, when allowing for 10 errors, we observed that the most stringent filtering method identified fewer variables than FEVS among those with smaller mean shifts (Figure 1(b)). Additional simulations were performed in which the sample size was increased from 5 to 20 samples per group, with the

Table I. Global null simulation results. Proportion of simulations with number of false positives greater than  $u$  (nominal value=0.05).

Method	$n=5$		$n=20$		$n=50$	
	$u=0$	$u=10$	$u=0$	$u=10$	$u=0$	$u=10$
FEVS	0.0461	0.0446	0.0481	0.0489	0.0495	0.0540
0 per cent (No filtering)	0.0463	0.0447	0.0555	0.0492	0.0515	0.0502
50 per cent	0.0463	0.0442	0.0503	0.0516	0.0512	0.0507
90 per cent	0.0483	0.0450	0.0459	0.0538	0.0471	0.0519
Longest list (Naïve filtering)	0.1766	0.2239	0.2082	0.2604	0.2149	0.2614

*Note:* The expression of 50 000 variables for two groups of 5, 20 and 50 samples each ( $n$ ) was simulated independently, as described in Section 3.  $u$  is the number of false discoveries allowed for in each analysis. FEVS is the filtering-enhanced variable selection method proposed in this paper. ‘0 per cent (No filtering)’ is the method in which data are not filtered; ‘50 per cent’ and ‘90 per cent’ refer, respectively, to the methods in which 50 per cent and 90 per cent of the variables are filtered out. ‘Longest list’ is the method in which a fixed-percentage filtering method is selected after considering 10 different fixed-percentage filtering methods (filtering out 0 per cent, 10 per cent, 20 per cent, ..., 90 per cent of the variables) and choosing the one that identifies the longest list of differentially expressed genes.

Table II. Alternative case simulation results. Average number of the 300 truly differentially expressed variables that are identified by the various procedures as differentially expressed and the proportion of simulations with the number of false positives greater than  $u$  (nominal value=0.05).

Method	$n=5$						$n=20$						$n=50$											
	Independent		Correlated		Independent		Correlated		Independent		Correlated		Independent		Correlated									
	$u=0$	$u=10$	$u=0$	$u=10$	$u=0$	$u=10$	$u=0$	$u=10$	$u=0$	$u=10$	$u=0$	$u=10$	$u=0$	$u=10$	$u=0$	$u=10$								
FEVS	43.865	138.91	46.55	136.13	235.91	263.52	236.91	261.23	282.18	292.40	282.08	291.75	0.0452	0.0426	0.0440	0.0432	0.0510	0.0229	0.0524	0.0282	0.0450	0.0197	0.0475	0.0245
0 per cent	13.63	98.32	15.06	97.97	233.54	263.75	234.08	263.02	281.90	291.96	281.63	291.37	0.0476	0.0475	0.0450	0.0491	0.0549	0.0456	0.0493	0.0523	0.0499	0.0440	0.0482	0.0428
50 per cent	15.27	105.10	16.82	104.65	235.09	262.22	235.51	262.15	277.49	286.41	277.20	286.29	0.0470	0.0489	0.0456	0.0463	0.0513	0.0446	0.0492	0.0450	0.0502	0.0437	0.0509	0.0458
90 per cent	27.90	136.43	30.46	134.63	214.65	225.72	213.33	226.68	233.14	236.75	232.85	235.00	0.0454	0.0441	0.0453	0.0459	0.0485	0.0390	0.0491	0.0427	0.0457	0.0375	0.0514	0.0388
97.5 per cent	45.71	132.20	48.93	129.28	169.50	173.03	169.49	171.56	179.63	180.89	179.66	180.56	0.0468	0.0376	0.0450	0.0441	0.0450	0.0078	0.0502	0.0312	0.0440	0.0193	0.0462	0.0298

Note: Expression profiles consisting of 50 000 variables for two groups of 5, 20 or 50 samples each ( $n$ ) were generated as described in Section 3; independently for columns marked 'independent' and with a block exchangeable correlation structure for columns marked 'correlated'.  $u$  is the number of false discoveries allowed for by each algorithm. The filtering methods are indicated with the same terminology used in Table I. For each pair of rows, the number in the top row is the average number of truly differentially expressed variables that was identified, and the number in the bottom row is the proportion of simulations in which the number of false positives exceeded  $u$ .



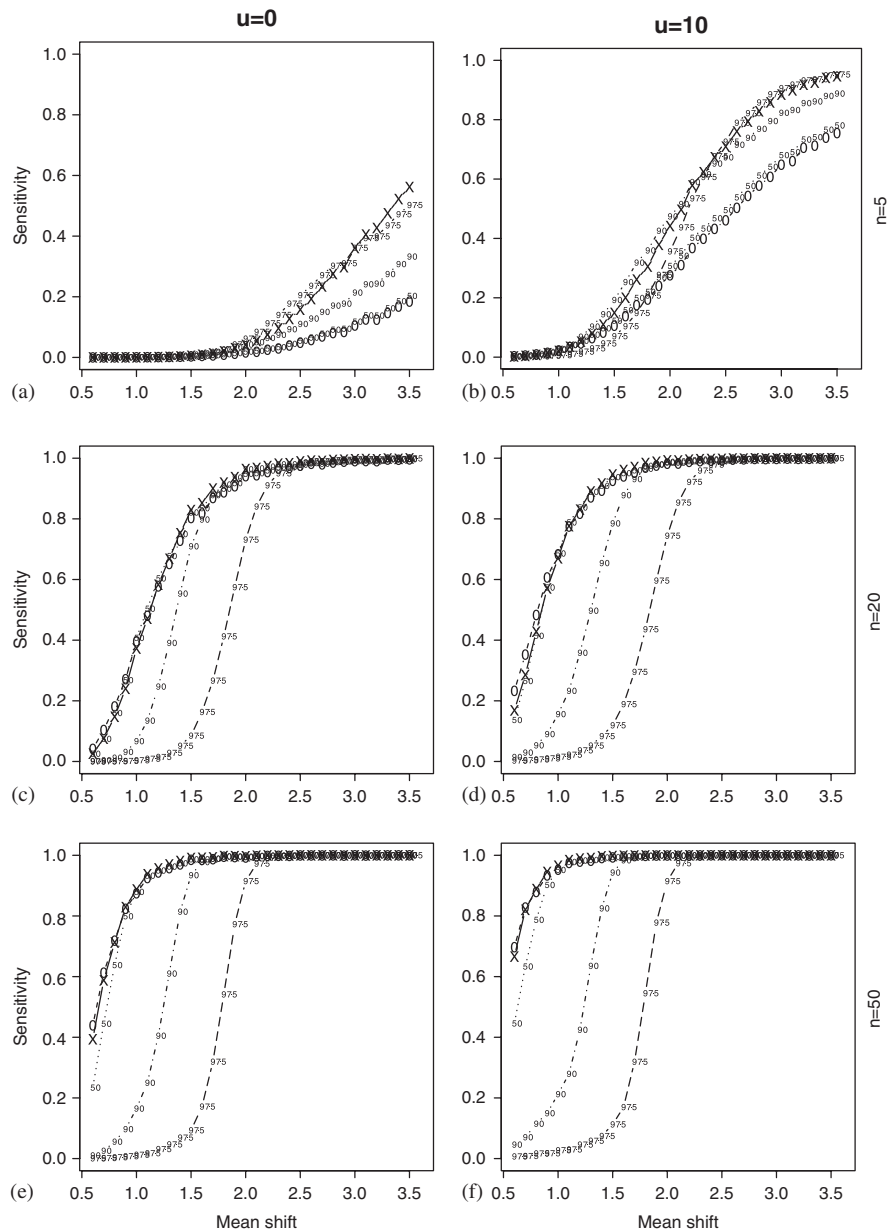


Figure 1. Proportion of truly differentially expressed genes, as a function of mean shift, identified by different filtering methods. Note:  $n$  is the number of samples per group and  $u$  is the number of false discoveries allowed for in each analysis. The simulation setting is the same as that described for Table II (for  $n=5$ , 20 and 50 sample per group), with all the variables being independent. The results indicated by a '0' are those obtained with no filtering, those indicated with an 'X' are those from the filtering-enhanced variable selection, while '50', '90' and '97.5' are those corresponding, respectively, to the 50 per cent, 90 per cent and 97.5 per cent fixed-percentage filtering.

other parameters maintained (first two columns of the second panel in Table II). When allowing for 0 and 10 errors, the results showed that filtering out variables did not increase the sensitivity for detecting differentially expressed variables and that the FEVS performance was very similar to no filtering, while fixed-90 and 97.5 per cent filtering failed to identify many of the variables with smaller mean shifts and did not identify more variables with high mean shifts (Table II and Figure 1(c) and (d)). This simulation shows that more stringent fixed-percentage filtering methods do not always identify the largest number of variables, and that the results obtained with FEVS are not always similar to those obtained with the fixed-97.5 per cent filtering. However, in both of these examples, FEVS was most sensitive or close to the most sensitive method. Very similar results were observed when the sample size was further increased to 50 samples per group (Table II, third panel and Figure 1(e) and (f)); in this simulation the sensitivity for identifying differentially expressed variables was close to 1 when no filtering was used (both for  $u=0$  and  $u=10$ ), and this high sensitivity was maintained by FEVS.

Comparable results were obtained when the variables were not all independent but simulated under a block exchangeable correlation structure, in which the variables in the same block were correlated while the variables from different blocks were independent. In particular, Table II (second two columns of each panel) displays the results in which blocks contained 100 correlated variables, pairwise correlation within each block was equal to 0.3 and the 300 differentially expressed variables were included in the first three blocks.

The simulation results shown so far, together with additional simulations (data not shown), suggest that FEVS is more useful in gaining sensitivity in situations where sample size is small ( $n=5$  per class), when compared with no filtering. We explain why this is true with class-independent filtering in the discussion. However, we note that one can identify examples in which FEVS is more sensitive than no filtering even with larger sample sizes. For example, for  $n=20$  per group, using an intraclass variance of 1 for all the 50 000 variables and simulating 100 of differentially expressed variables with a mean shift equal to 1.92, the average sensitivity of FEVS was 74.3 per cent, compared with the 58.9 per cent sensitivity obtained without filtering variables. The next section discusses other examples in which FEVS has higher overall sensitivity than any of its constituent filtering methods.

*3.1.3. Examples when FEVS outperforms all of its constituent filtering methods.* To help demonstrate that FEVS can potentially perform better than any of its constituent filtering methods, we considered a further situation with 2 sets of 50 differentially expressed variables out of the 50 000 variables. For one set of variables, the univariate tests have high power for identifying differentially expressed features due to small intraclass variability. The second set has a larger difference of means between the two classes, but lower power because of larger intraclass variability. The mean shifts were chosen to obtain 95 per cent univariate-power based on Bonferroni adjustment for the first set of variables on the complete data set (without filtering) and 80 per cent power for the second set if only 5000 variables were retained in the data set (90 per cent filtering).

Results of a simulation with  $n=5$  per group are presented in the left-most panel of Table III. For this simulation, the mean shift between the two classes for the differentially expressed variables in set 1 was 1.18 and the intraclass standard deviation ( $\sigma$ ) was fixed to 0.1; in the second set of differentially expressed variables, the mean shift was 7.37 and the intraclass standard deviation was 1. The number of allowed errors was  $u=0$ ; the null variables were simulated into two equal size groups, with one group having the same intraclass variance as the group 1 differentially expressed variables and the other group having the same intraclass variances as group 2 (24 950

Table III. Numbers of the 100 truly differentially expressed variables identified as differentially expressed.

Method	$n=5$			$n=20$			$n=50$		
	Total	Identified from set 1	Identified from set 2	Total	Identified from set 1	Identified from set 2	Total	Identified from set 1	Identified from set 2
FEVS	77.78	34.25	43.52	84.19	44.94	39.25	80.01	45.38	34.63
0 per cent	62.58	45.46	17.12	77.26	47.86	29.40	78.25	47.54	30.71
50 per cent	70.61	47.71	22.90	32.92	0.00	32.92	33.54	0.00	33.54
90 per cent	28.02	0.00	28.02	37.01	0.00	37.01	28.08	0.00	28.08
97.5 per cent	37.28	0.00	37.28	35.79	0.00	35.79	17.90	0.00	17.90

*Note:* Two sets of 50 differentially expressed variables out of 50 000 were simulated independently for 2 groups of 5, 20 and 50 samples each ( $n$ ), as described in Section 3. The first set contained variables with small mean shifts and small intraclass variability, while the second set contained variables with bigger mean shifts and bigger intraclass variability. All methods were controlling for  $\mu=0$  errors. The filtering methods are indicated with the same terminology used in Table I.

with  $\sigma=0.1$  and 24 950 with  $\sigma=1$ ), and all the variables were independent. The results reported in Table III (first panel) show that in this particular setting FEVS was not similar to any of the fixed-percentage filtering methods, none of which reached a higher sensitivity to detect differentially expressed variables than FEVS. In particular, FEVS identified about 87 per cent of the variables from the second set, with the more stringent fixed-percentage filtering methods doing slightly worse (75 per cent). On the other hand, FEVS identified about 68 per cent of the variables from the first set, all of which were filtered out by the more stringent methods. A very similar result was obtained when we considered  $n=20$  samples in each class, and simulated mean shifts of 0.25 and 1.91 for the two differentially expressed sets, with all the other parameters kept equal to those described for the simulation with  $n=5$  (see Table III, second panel). For  $n=50$ , where we simulated mean shifts of 0.14 and 1.11, the sensitivity for the first set was 95 per cent for no filtering and 90 per cent for FEVS, while for the second set it was higher for FEVS (69 per cent vs 61 per cent, see Table III, third panel).

### 3.2. An application to microarray data from breast cancer

Sotiriou *et al.* [15] analyzed cDNA gene expression profiles from 99 tumor specimens from breast cancer patients. In addition to gene expression values for 7650 genes (clones) pre-processed as described in [15], there was standard prognostic variable information available for each patient. (The data are publicly available at <http://www.pnas.org/cgi/content/full/1732912100/DC1> in Supporting Tables II and III.) Using additional information publicly available on the NCI Microarray Database (mAdb) website (<http://nciarray.nci.nih.gov/cgi-bin/gipo>) for the array print set used in this study (Hs-ATC7.6k-v5p4-020801), we identified 292 spots on the array for which clones failed the sequence verification and changed their identity. The annotation of the remaining clones was updated by submitting the IMAGE clone IDs to Source (<http://source.stanford.edu>). Updates from all of these databases were downloaded on 31 July 2007.

Here we consider two two-class comparisons based on two-sample *t*-tests and control for the number of false positives with 95 per cent confidence. For each comparison, we restricted attention to genes for which the number of missing values was less than the number of specimens in the class with fewer observations minus 2. We use FEVS based on the interquartile range ranking of the genes and 9999 resampled permutations.

The first comparison is for patients with grade 1 or 2 tumors ( $n=54$ ) vs patients with grade 3 tumors ( $m=45$ ) with  $k=7498$  genes. Allowing for no errors ( $u=0$ ), the MPT-based procedure without any filtering of the genes identifies 6 genes, while FEVS identifies 20 genes. Allowing for 10 errors ( $u=10$ ), 94 and 124 genes are identified by the no-filtering and the FEVS procedures, respectively (Table IV).

The second comparison is for patients with estrogen receptor (ER) negative status ( $n=34$ ) vs patients with ER positive status ( $m=65$ ) with  $k=7470$  genes (ER measured with ligand-binding assay). Allowing for no errors ( $u=0$ ), the MPT-based procedure with no filtering identifies 172 genes and FEVS identifies 199 genes. Allowing for 10 errors ( $u=10$ ), the number of genes identified without filtering out any of the genes is higher than that obtained with FEVS (503 and 472 genes, respectively, with 424 genes included in both the lists, Table IV). Most of the genes not identified by FEVS had small gene expression differences between the two groups (data not shown).

Even though in the second comparison the number of genes identified decreases as more genes are filtered out, the genes included in the lists obtained with more stringent filtering methods were not

Table IV. Number of genes identified as differentially expressed using various filtering strategies, allowing for  $u$  false positives for two comparisons involving breast cancer specimens of Sotiriou *et al.* [15].

Method	Tumor grade (1 or 2 vs 3)		Tumor ER status (positive vs negative)	
	$u=0$	$u=10$	$u=0$	$u=10$
FEVS	20	124	199	472
0 per cent	6	94	172	<b>503</b>
10 per cent	7	99	174	501
20 per cent	8	101	177	502
30 per cent	9	106	180	494
40 per cent	10	106	<b>186</b>	483
50 per cent	13	114	180	466
60 per cent	14	<b>115</b>	179	443
70 per cent	13	110	172	414
80 per cent	14	106	160	350
90 per cent	<b>19</b>	98	136	247

*Note:* The fixed-percentage filtering methods with which the highest numbers of genes were identified are indicated in bold. The same terminology of Table I is used to indicate the filtering methods.

necessarily included in the list obtained without filtering data. For example, just 71 per cent of the 247 genes obtained with a fixed-90 per cent filtering were included in the 503 genes obtained with no filtering. As was observed with the simulations, the fixed-percentage filtering method with which the highest number of differentially expressed genes was identified varied, depending on the number of errors allowed for and, in this case, also on the class-defining variable being analyzed (Table IV). We evaluated the biological plausibility of the genes identified by FEVS but not by the MPT-based procedure without any filtering of the genes (FEVS-exclusive genes), checking whether these genes were previously identified by other investigators using microarray technology. We used the breast cancer data sets included in Oncomine [16] as of July 2007 (<http://www.oncomine.org>) for which the ER (19 data sets) or the grade information (15 data sets) was available and evaluated the differential expression using the statistical significance test provided in Oncomine ( $t$ -test for gene expression classified by the ER status and correlation between gene expression and tumor grade (1, 2, or 3)). The list of selected studies and the results of this analysis are included in the Supplementary Information, available at <http://linus.nci.nih.gov/Data/LusaL/bioinfo/>. Results showed that most of the FEVS-exclusive genes were found differentially expressed also in independent data sets. For grade, allowing for 0 errors (10 errors) we identified 14 (32) unique FEVS-exclusive genes, most of which were identified by Oncomine as highly significantly associated with grade in at least one of the independent data sets (Table S1). For the ER status, allowing for 0 errors (10 errors) we identified 29 (31) unique known FEVS-exclusive genes, most of which were identified by Oncomine as highly significantly associated with the ER status in at least one of the independent data sets (Table S2).

#### 4. DISCUSSION

In this paper we presented a new approach (FEVS) to the identification of the differentially expressed variables for high-dimensional data sets, which can be used in any situation in which

multivariate permutation methods are applicable. The approach combines the results obtained after applying a variety of variable filtering methods. The aim of this method is to diminish the multiple comparisons problem, while avoiding the arbitrariness of the choice of a pre-specified filtering method.

We showed with a limited set of simulations and an application that it does not seem feasible to identify a universally optimal filtering method, i.e. a single filtering strategy that helps to identify the largest number of truly differentially expressed variables under all circumstances. Also, we showed with simulated and real data that the sets of variables identified by applying filtering methods with different stringency are not necessarily heavily overlapping; therefore, indicating the possible advantage of using a method that combines the results obtained from multiple analyses utilizing different filtering strategies over a method that attempts to identify the single best filtering method.

Simulation results showed that even though FEVS could be more sensitive than no filtering for moderate and large sample sizes, the greatest gains in terms of sensitivity were obtained by FEVS when the sample size was small. This is to be expected with class-independent filtering based on total variation when the intraclass variation is the same for null and differentially expressed variables: mean effects large enough to be affected by the filtering will have, with large sample sizes, such high power that they would be identified even if there was no filtering. On the other hand, if differentially expressed variables have larger intraclass variability than the null hypothesis variables, then filtering and FEVS would identify more of these variables even with large sample sizes. It is interesting to note that using a large data set of breast cancer samples and comparing low to high grade tumors or ER status, FEVS identified a larger number of differentially expressed genes compared with no filtering, most of which were previously identified by other microarray studies as being related to tumor grade or ER status, respectively.

In this paper we restricted our attention to two-class comparisons in the presentation of the method and both in the simulations and in the example using real data. However, FEVS is based on a multivariate permutation procedure [12] and can therefore be used in all the situations in which multivariate permutation methods are applicable, which include, besides unpaired or paired two-group comparisons, also K-group comparisons, linear or logistic regression with one independent variable, and survival analysis.

Any class-independent filtering ranking-statistic can be used to rank the variables when using FEVS. We used the interquartile range in all of the class-independent FEVS examples, which was previously suggested as a good choice [17]. We also considered two other filtering ranking-statistics, the variance and the 95th minus 5th percentile, obtaining very similar results to those presented for the simulated and real data (results not shown). In principle, the FEVS method could be used to combine filtering methods based on different filtering ranking-statistics. We did not explore the performance of such a strategy, which would be more cumbersome from a computational point of view. In addition, it is not obvious that it would prove to be useful in practice given the similarity of the results that we observed using different filtering ranking-statistics.

An approach with some relationship to FEVS is DEDS (differential expression via distance synthesis, [18]). This approach was proposed to synthesize different statistics that measure the same quantity of interest, controlling the false discovery rate with a permutation-based algorithm. The idea behind DEDS is to identify the variables that rank high according to all statistics; therefore using the concept of an 'intersection'. FEVS looks instead for the 'union' of the variables identified by different filtering methods. In principle, FEVS also could be used to combine the results obtained using different test statistics instead of using different filtering methods.

Other approaches that have some similarity with FEVS are the data-driven weighted procedures for the control of the false discovery rate (expected value of the proportion of false discoveries) [19] and the therein reviewed (data-driven) weighted methods for familywise error control. In [19] the proposed weights are the total variances of the variables and the procedures rely on univariate  $p$ -values rather than using multivariate permutation-based methods. Similar to our simulation results, they report better performance of their method in terms of sensitivity when the sample size is small.

Others have considered class-dependent methods for filtering [20, 21]. We also explored class-dependent filtering methods with FEVS. Note that with class-dependent filtering, the equality in equation (2) would no longer hold, and would be expected to be the inequality  $\leq$ . This would guarantee the correct error rates for FEVS, but would reduce the sensitivity of the method. Simulations (not shown) verified that there was no clear benefit from the use of class-dependent filtering methods; when compared with a class-independent FEVS, neither a larger number of variables were identified, nor did the variables included in the lists derived with the class-dependent filtering exhibit a larger mean difference between classes.

We presented the FEVS method limited to the case in which the number of false discoveries is controlled, but the method can be extended to control approximately the proportion of false discoveries, modifying the algorithm proposed by Korn *et al.* [12] in a way similar to what was proposed in this paper for the control of the number of false positives [22]. One would expect gains in the numbers of identified variables as was seen for the results presented in this paper for controlling the number of false discoveries.

#### ACKNOWLEDGEMENTS

We gratefully acknowledge the funding provided to L. L. by an Italy–U.S.A. Fellowship of the Istituto Superiore di Sanità on Oncological Pharmacogenomics—Seroproteomics and AIRC (Associazione Italiana per la Ricerca sul Cancro, individual grant to Marco A. Pierotti and Manuela Gariboldi). This study utilized the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health, Bethesda, MD.

#### REFERENCES

1. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002; **12**:111–140.
2. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. *Design and Analysis of DNA Microarray Investigations*, Chapter 7. Springer: New York, 2004.
3. Dudoit S, van der Laan MJ. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics, Chapters 1–3. Springer: New York, 2008.
4. Yamanaka R, Arao T, Yajima N, Tsuchiya N, Homma J, Tanaka R, Sano M, Oide A, Sekijima M, Nishio K. Identification of expressed genes characterizing long-term survival in malignant glioma patients. *Oncogene* 2006; **25**:5994–6002.
5. Fält S, Holmberg K, Lambert B, Wennborg A. Long-term global gene expression patterns in irradiated human lymphocytes. *Carcinogenesis* 2003; **24**:1837–1845.
6. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, O’Connell P. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet* 2003; **362**:362–369.
7. Debernardi S, Lillington DM, Chaplin T, Tomlinson S, Amess J, Rohatiner A, Lister TA, Young BD. Genome-wide analysis of acute myeloid leukemia with normal karyotype reveals a unique pattern of homeobox gene expression distinct from those with translocation-mediated fusion events. *Genes Chromosomes Cancer* 2003; **37**:149–158.

8. Ma X-J, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT, Tran Y, Tran D, Tassin A, Amon P, Wang W, Wang W, Enright E, Stecker K, Estepa-Sabal E, Smith B, Younger J, Balis U, Michaelson J, Bhan A, Habin K, Baer TM, Brugge J, Haber DA, Erlander MG, Sgroi DC. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004; **5**:607–616.
9. Whistler T, Jones JF, Unger ER, Vernon SD. Exercise responsive genes measured in peripheral blood of women with chronic fatigue syndrome and matched control subjects. *BMC Physiology* 2005; **5**:5. DOI: 10.1186/1472-6793-5-5.
10. Bonome T, Lee JY, Park DC, Radonovich M, Pise-Masison C, Brady J, Gardner GJ, Hao K, Wong WH, Barrett JC, Lu KH, Sood AK, Gershenson DM, Mok SC, Birrer MJ. Expression profiling of serous low malignant potential, low-grade, and high-grade tumors of the ovary. *Cancer Research* 2006; **65**:10602–10612.
11. Argos M, Kibriya MG, Parvez F, Jasmine F, Rakibuz-Zaman M, Ahsan H. Gene expression profiles in peripheral lymphocytes by arsenic exposure and skin lesion status in a Bangladeshi population. *Cancer Epidemiology Biomarkers and Prevention* 2006; **15**:1367–1375.
12. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 2004; **124**:379–398.
13. David HA. *Order Statistics* (2nd edn). Wiley: New York, 1981; 15.
14. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 2003; **19**:2448–2455.
15. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America* 2003; **100**:10393–10398.
16. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 2004; **6**:1–6.
17. Von Heydebreck A, Huber W, Gentleman R. Differential expression with the bioconductor project. *Encyclopedia of Genomics, Proteomics and Bioinformatics*. Wiley: New York, 2004.
18. Yang YH, Xiao Y, Segal MR. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics* 2005; **21**:1084–1093.
19. Finos L, Salmaso L. FDR- and FWE-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference* 2007; **137**:3859–3870.
20. Hero AO, Fleury G, Mears AJ, Swaroop A. Multicriteria gene screening for analysis of differential expression with DNA microarrays. *EURASIP Journal on Applied Signal Processing* 2004; **1**:43–52.
21. van de Wiel MA, Kim KI. Estimating the false discovery rate using nonparametric deconvolution. *Biometrics* 2007; **63**:806–815.
22. Korn EL, Li M-C, McShane LM, Simon R. An investigation of two multivariate permutation methods for controlling the false discovery proportion. *Statistics in Medicine* 2007; **26**:4428–4440.