

Methodology article

Open Access

## Dissecting systems-wide data using mixture models: application to identify affected cellular processes

J Peter Svensson<sup>1,2</sup>, Renée X de Menezes<sup>3</sup>, Ingela Turesson<sup>2</sup>,  
Micheline Giphart-Gassler<sup>1</sup> and Harry Vrieling\*<sup>1</sup>

Address: <sup>1</sup>Department of Toxicogenetics, Leiden University Medical Centre, P.O. Box 9503, 2300 RA Leiden, the Netherlands, <sup>2</sup>Department of Oncology, Radiology and Clinical Immunology, Academic Hospital, 751 85 Uppsala, Sweden and <sup>3</sup>Department of Medical Statistics, Leiden University Medical Centre, P.O. Box 9604, 2300 RA Leiden, the Netherlands

Email: J Peter Svensson - p.svensson@lumc.nl; Renée X de Menezes - r.x.menezes@lumc.nl; Ingela Turesson - ingela.turesson@onkologi.uu.se; Micheline Giphart-Gassler - m.giphart-gassler@lumc.nl; Harry Vrieling\* - h.vrieling@lumc.nl

\* Corresponding author

Published: 14 July 2005

Received: 08 December 2004

BMC Bioinformatics 2005, 6:177 doi:10.1186/1471-2105-6-177

Accepted: 14 July 2005

This article is available from: <http://www.biomedcentral.com/1471-2105/6/177>

© 2005 Svensson et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Functional analysis of data from genome-scale experiments, such as microarrays, requires an extensive selection of differentially expressed genes. Under many conditions, the proportion of differentially expressed genes is considerable, making the selection criteria a balance between the inclusion of false positives and the exclusion of false negatives.

**Results:** We developed an analytical method to determine a *p*-value threshold from a microarray experiment that is dependent on the quality and design of the data set. To this aim, populations of *p*-values are modeled as mathematical functions in which the parameters to describe these functions are estimated in an unsupervised manner. The strength of the method is exemplified by its application to a published gene expression data set of sporadic and familial breast tumors with *BRCA1* or *BRCA2* mutations.

**Conclusion:** We present an objective and unsupervised way to set thresholds adapted to the quality and design of the experiment. The resulting mathematical description of the data sets of genome-scale experiments enables a probabilistic approach in systems biology.

### Background

Functional analysis of microarray data, e.g. to reveal enrichment of promoter sequences, metabolic pathways or signalling cascades, requires an extensive selection of differentially expressed genes. Arbitrarily chosen thresholds for fold-changes and/or significance of change are commonly used to split the genes in subsets of alternative (differentially expressed) and null (non-differentially expressed) genes, with acceptable proportions of false positives and false negatives. Which proportion is acceptable depends on the research question. For example, to

find markers of a treatment or signatures of a mutation, it is usually enough to fix the FDR-control level [1], and select the most significant alternative genes. However, most differentially expressed genes will not be selected. For other purposes, e.g. when comparing differences between treatments or mutations, there is need for more thorough comparisons. For each gene we would like to know the probability of it being alternative or null. Such an approach will let us select genes that are truly – or truly not – differentially expressed, with associated estimates of false positives.

The differential expression of genes is evaluated after a comparison of gene expression levels, yielding a list of  $p$ -values. Commonly, a threshold is then fixed for either the number of genes selected, or for their  $p$ -value, and those genes corresponding to the  $k$  smallest  $p$ -values are classified as alternative. However, because of the large number of  $p$ -values involved, even if all features are null some small  $p$ -values may be observed due to pure chance. Thus, it is important to consider an estimate for the proportion of alternative features while determining the threshold.

The problem of estimating the proportion of alternative and null features has been handled by several authors [2-5]. In many experimental settings the alternative features make up a considerable fraction. This is a consequence of genes being connected in networks; altered expression of one gene can affect expression levels of multiple targets. Begley and co-workers [6] showed that the methylating agent MMS induced changes in the transcription level of 33% of the genes in the *S. cerevisiae* genome and oxidizing agent t-BuOOH altered 38% of the genes. In a study of familial and sporadic human breast cancers by Hedenfalk and co-workers [7], Storey and Tibshirani [5] estimated 33% of the genes to be differentially expressed between *BRCA1* and *BRCA2* mutation positive tumors. However, even though over 1,000 genes were considered to be changed, only 160 genes could be selected with a pFDR of 5% as a consequence of the overlap between alternative and null genes. It is intuitive that using only 16% of the differentially expressed genes for functional analysis is suboptimal, and we will show that this is indeed the case. To detect subtle but coordinated changes in gene expressions, it is advantageous to examine the joint behavior of inter-connected sets of genes [8], which implies the need for a generous selection. Significant changes in cellular processes can be detected in cases where the alteration of individual gene expressions is not significant [9].

The threshold determination has so far been left to the researcher. In practice it is either guided by limitations in the number of features to be verified after being classified as alternative, or by the desired false positive proportion in the final list. However, with the increasing rate at which new experiments are being performed, and decreasing cost of verification experiments, there will be a growing need for unsupervised ways of determining thresholds yielding a list of features with acceptable false positive and false negative rates.

We present here an objective and unsupervised way to set thresholds that are adapted to the quality and design of the experiment. Description of the distribution of both null and alternative  $p$ -values as mathematical functions will give the researcher the possibility to select genes depending on the probability of being alternative or null.

## Result

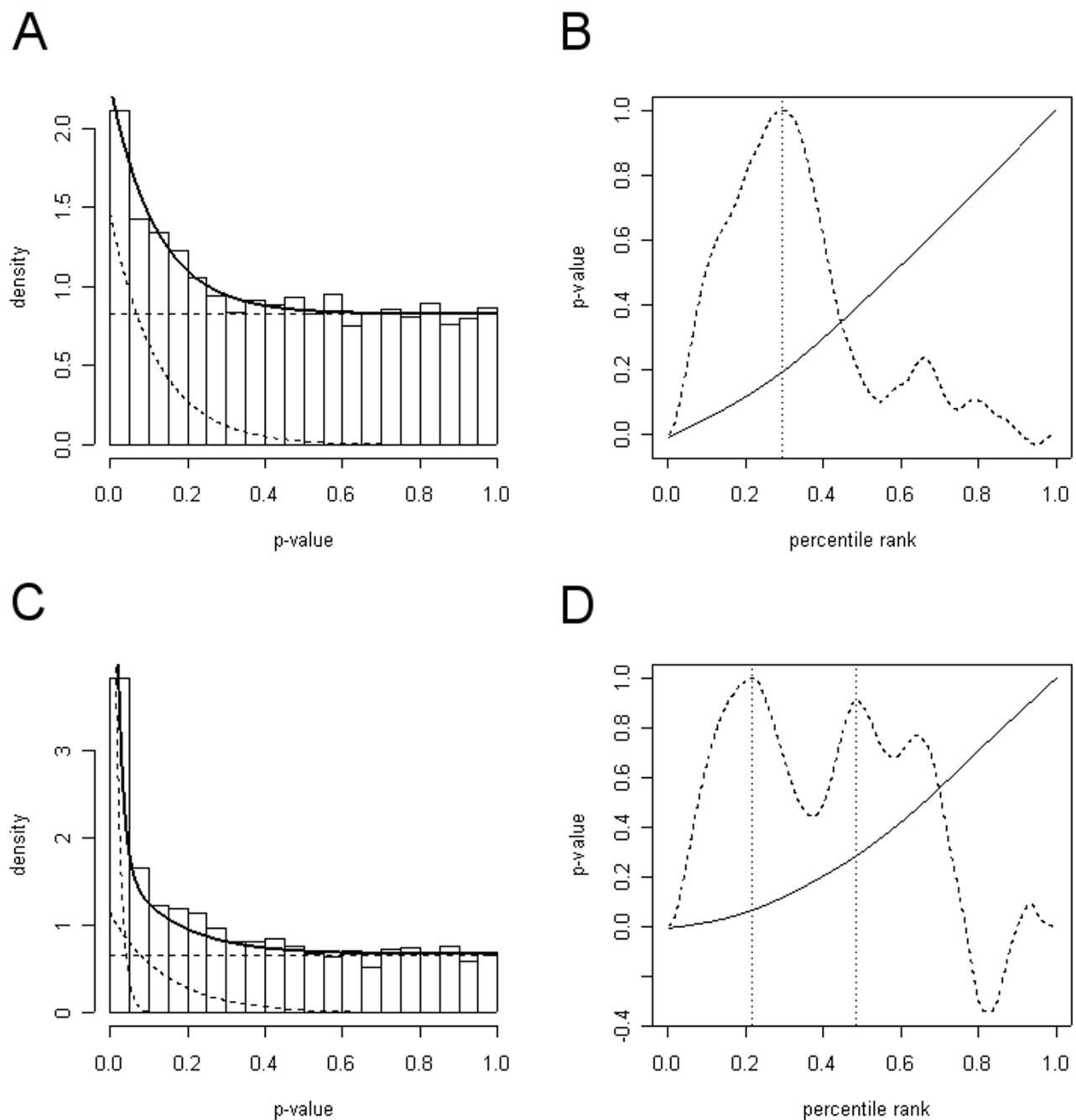
### A case study

Hedenfalk et al [7] determined gene expression patterns in 21 tumors from breast cancer patients. Seven tumors were sporadic with unknown mutations, whereas the remainder were familial cancers, in which one of the known genes associated with breast cancer was mutated: *BRCA1* (7 patients) or *BRCA2* (8 patients). RNAs from the three different sources were hybridized to arrays containing 6,512 cDNA clones. After discarding low quality spots [5,7], more than 3,000 clones remained. The measurements from these genes were tested against the null hypothesis that there is no differential expression across two conditions [5]. The resulting  $p$ -values can be visualized in density histograms (figure 1A,C and supplementary material). The  $p$ -values appear to follow the expected distribution as being composed of a population of truly null (non-differentially expressed) genes with  $p$ -values uniformly distributed among  $[0, 1]$  and a population of truly alternative (differentially expressed) genes with  $p$ -values that tend to be close to zero. Similar shapes of distributions can be generated for all four tested comparisons, *BRCA1* to sporadic, *BRCA2* to sporadic, *BRCA1* to *BRCA2* and both *BRCA1* and *BRCA2* to sporadic.

### Decomposing distributions and setting thresholds

To further examine the gene expression responses, we describe the alternative and null distributions as mathematical functions. For this purpose, we plot the genes sorted by their  $p$ -values,  $p$ , against the percentile ranks (rank position/the total number of  $p$ -values),  $y$  (figure 1B) to reveal  $F^{-1}$ , the inverse of the underlying cumulative density function.  $F^{-1}$  is estimated non-parametrically. When comparing *BRCA1* mutation-positive tumors against sporadic tumors,  $\hat{F}^{-1}$  is convex upward on the entire interval

$y \in [0, 1]$ . The curvature  $(\frac{d^2}{dy^2} F^{-1}(y))$  (dotted line in figure 1B) has a major peak (at  $y_0 = 0.30$  corresponding to  $p_0 = 0.19$ ) indicating that there is one dominating population of alternative genes. The expression for  $f(p_0)$  gives us a rough  $\hat{\pi}_0$  estimated to 0.77. Other methods specifically designed to estimate  $\hat{\pi}_0$  yielded similar results; Storey and Tibshirani's method [5] led to  $\hat{\pi}_0 = 0.83$  and the method developed by Schweder and Spjøtvoll [4] resulted in  $\hat{\pi}_0 = 0.82$ . Under the assumption that the  $p$ -values of the alternative genes follow an exponential distribution, we calculate  $\hat{\lambda}$  to 8.3 (see methods section). With the parameters  $\hat{\lambda}$  and  $\hat{\pi}_0$  we can describe the null and alternative density functions (dashed lines in figure 1A). To select differentially expressed genes, we use the critical

**Figure 1**

**Distributions of *p*-values from the Hedenfalk data set.** (A,C) Density histograms for the more than 3,000 genes comparing (A) *BRCA1*-mutated and sporadic tumors or (C) *BRCA1*- and *BRCA2*-mutated tumors. The dashed lines show the estimated distributions of the alternative (curves) and null (horizontal lines) genes, and solid lines show the sum the estimated distributions. (B,D) Scatterplots where the *p*-value of each gene is plotted against the percentile rank. A smooth function is fitted (solid line) and the local maxima of the curvature (dashed line) are used to split the null and alternative populations of genes for the comparisons (B) *BRCA1*-mutated and sporadic tumors or (D) *BRCA1*- and *BRCA2*-mutated tumors.

point  $p_0$  as a  $p$ -value threshold. As a result, the null hypothesis was accepted for 2,083 genes with a  $p > 0.19$ . These genes all have at least a twice larger probability of being null than alternative. In general, since we have a mathematical description of  $f(p)$  any point can be chosen as threshold. For instance, we can calculate a point,  $p_1$ , where genes have at least a two times larger probability of being alternative than null. Hundred-and-eighty genes have a  $p$ -value below this threshold of  $p_1 = 0.04$ , and are selected as alternative. The proportion of false positives, as estimated by the ratio of integrals of  $f_0$  and  $f$  in the interval  $[0, p_1]$ , amounts to 20%. The pFDR estimated by Storey and Tibshirani [5] is 34%. Notably, in this comparison at a pFDR of 5%, being the proposed threshold by Storey and Tibshirani [5], no gene is selected as being differentially expressed.

Comparing expression data from *BRCA2* and sporadic tumors shows similar patterns ( $\hat{\pi}_0 = 0.75$  and  $\hat{\lambda} = 11.1$ ) (supplementary figure 1). Here, using  $p_0$  as the  $p$ -value threshold 2,168 genes are selected as being null. Only 20% of the genes with  $p < p_{0(BRCA1)}$  or  $p < p_{0(BRCA2)}$  are in common in both comparisons, suggesting that *BRCA1* and *BRCA2* affect different target genes.

Information on the genes affected in both *BRCA1* and *BRCA2* mutation-positive tumors can be obtained by comparing the data from the familial tumors together against the sporadic. The curvature of  $\hat{F}^{-1}$ , revealing the genes in common between *BRCA1* and *BRCA2* mutated tumors, has two local maxima (supplementary figure 1). Also when the data from *BRCA1* is compared to that from *BRCA2* mutated tumors (figure 1C–D), the curvature plot has several local maxima. These observations indicate that the original distributions might be constituted of  $p$ -values from two or more populations of truly alternative genes in addition to the truly null genes. Each of these groups is represented by a close to linear part in the plot of  $\hat{F}^{-1}(y)$ .

The critical points of the curvature of  $\hat{F}^{-1}(y)$  and the corresponding  $p$ -values let us estimate the parameters to describe the distributions of the groups.

### Simulation study

In order to verify how the method works in a case for which the result is known, a simulation study was performed. It was assumed for simplicity that each alternative  $p$ -value followed an exponential distribution with rate  $\lambda$ . For values of  $\lambda$  between 1 and 100, and varying proportions  $\pi_0$  of alternative  $p$ -values within  $[0, 1]$ , a set of 10,000 independent  $p$ -values was generated. For each of these sets, an estimate  $\hat{\pi}_0$  was computed for  $\pi_0$  using the

method developed by Storey and Tibshirani [5]. Subsequently, an estimate  $\hat{\lambda}$  for  $\lambda$  was determined via a non-parametric approach (see methods section). A graph of the estimate  $\hat{\lambda}$  for all combinations of  $\lambda$  and  $\pi_0$  values, is shown in figure 2. The relative error in  $\hat{\lambda}$  was determined as  $\frac{(\lambda - \hat{\lambda})^2}{\lambda^2}$  and the proportion of false positives and false negatives were calculated by  $\int_0^{p_0} f_0 / \int_0^{p_0} f$  and  $\int_{p_0}^1 f_1 / \int_{p_0}^1 f$  respectively.

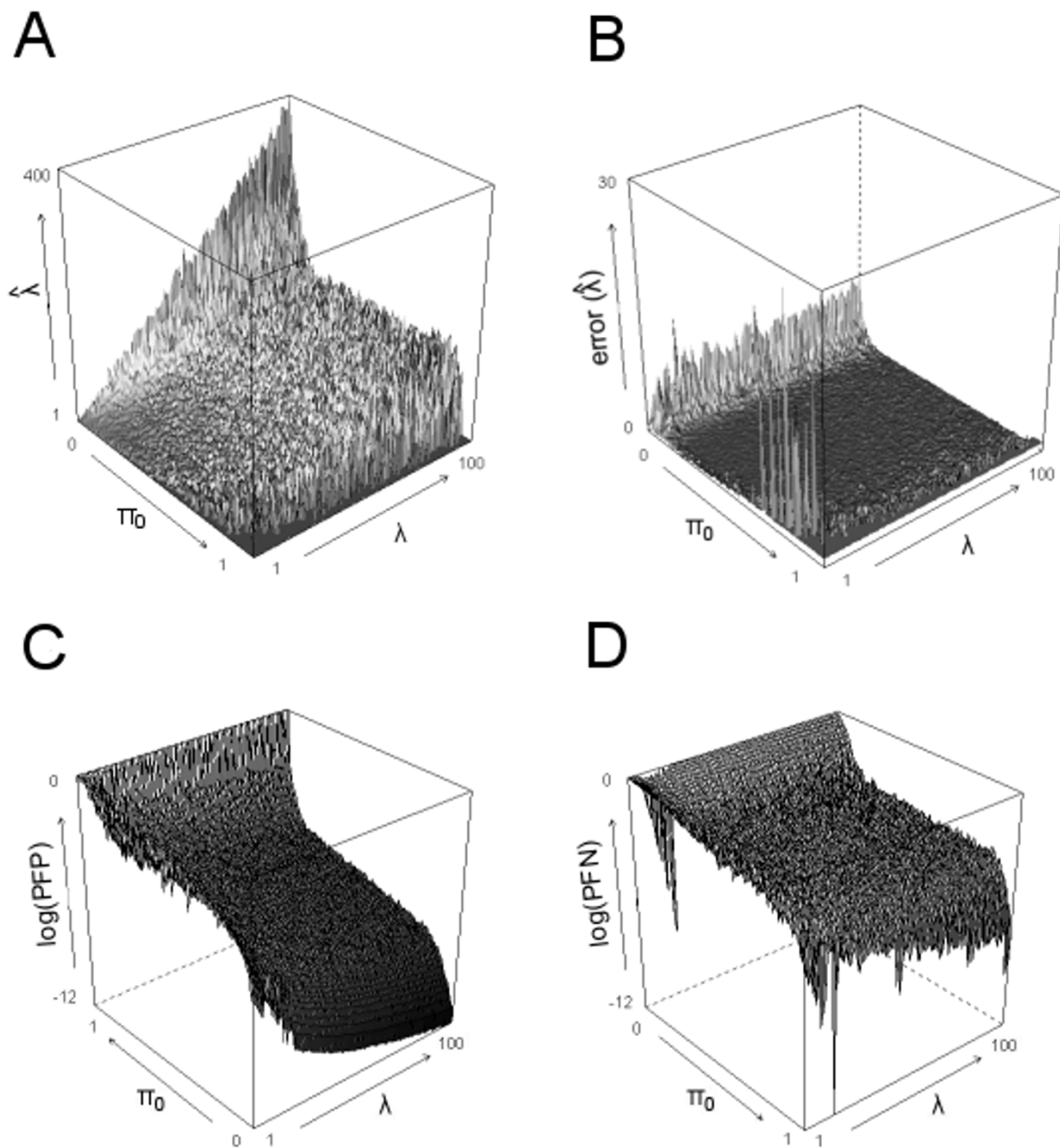
The results show that  $\hat{\lambda}$  gives a reasonable estimate of  $\lambda$  for  $0.20 < \pi_0 < 0.95$  and  $\lambda > 5$  (figure 2). The proportion of false positives obtained for these values is  $0.03 \pm 0.01$  (mean  $\pm$  s.d.), and the proportion of false negatives is  $0.04 \pm 0.03$ . When few alternative features are present ( $\pi_0 = 0.95$ ),  $\lambda$  cannot be determined. Also when the alternative distribution is close to uniform ( $\lambda < 5$ ),  $\hat{\lambda}$  fails to correctly estimate  $\lambda$ . For  $\pi_0 = 0.20$ ,  $\lambda$  is over-estimated and the method yields a conservative threshold, meaning that the proportion of false positives is low on the expense of a high proportion of false negatives. However,  $\pi_0$  being  $\leq 0.20$  is of relatively little practical interest, as typically by design, studies already contain a large fraction of unchanged genes.

We also compared the proportions of false positives and false negatives between different methods for selecting genes, e.g. while controlling the family wise error-rate (FWER) or the FDR (figure 3). The proportion of false calls largely depends on the separation between alternative and null features (the value of  $\lambda$ ) for thresholds based directly on the  $p$ -values. As to be expected, when correcting for multiple testing, i.e. controlling the FWER or FDR, the proportion of false positives remains constant.

Similar calculations can be made for complex populations with more than one population of alternative features.

### Functional analysis

To discern alternated cellular processes, we performed an analysis at gene set level [8,9]. To obtain a description of the function of the affected genes ( $p < p_0$ ) in the BRCA study, the genes on the arrays were associated with Gene Ontology (GO) terms [10]. The annotation of the genes in the original data files were updated to the Unigene 170 build (24 April 2004). Of the informative transcripts on the array, 62% could be attributed to gene sets that represent specific GO terms. Subsequently, the hypergeometric

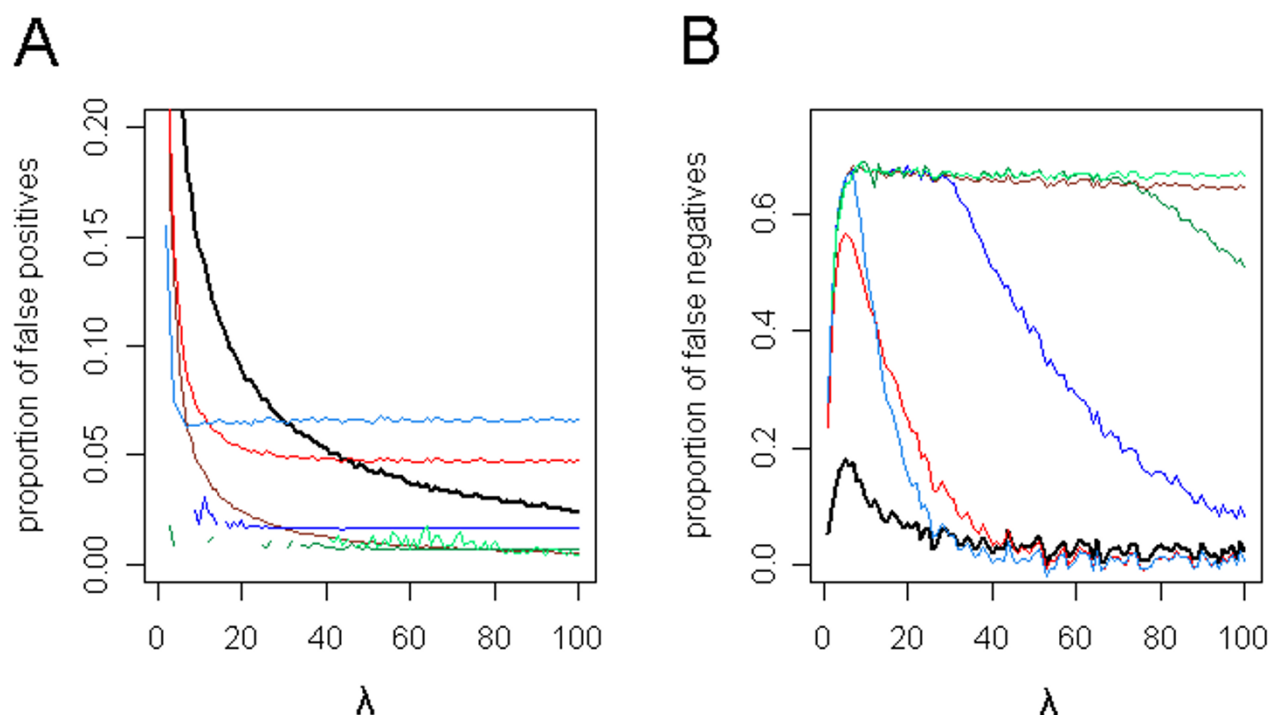
**Figure 2**

**Calculation of  $p_0$  and estimations of  $\lambda$  from a simulation study.** 10,000  $p$ -values were simulated in a mixture of an

alternative and a null distribution. As functions of  $\lambda \in [1, 100]$  and  $\pi_0 \in [0, 1]$  are shown: (A)  $\hat{\lambda}$ , (B) the relative error in

$\hat{\lambda} (= \frac{(\hat{\lambda} - \lambda)^2}{\lambda^2})$ , (C) the proportion of false positives (PFP) and (D) false negatives (PFN) when selecting the genes with  $p < p_0$ .

NB the inverse arrow of  $\pi_0$  in (C).

**Figure 3**

**Comparisons between  $p_0$  and other selection methods in a simulation study.** 10,000  $p$ -values were simulated in a mixture of an alternative and a null distribution. The proportions of false positives (**A**) and false negatives (**B**) are depicted as functions of  $\lambda \in [1, 100]$ ,  $p_0$  is set at 0.67. Selecting the genes with  $p < p_0$  (black) is compared to: selecting the genes with  $p < 0.001$  (brown) or  $p < 0.1$  (red); controlling the Benjamini-Hochberg FDR at  $\alpha = 0.05$  (dark blue) or  $\alpha = 0.2$  (light blue); the Benjamini-Yekutieli FDR at  $\alpha = 0.2$  (dark green) and the FWER at  $\alpha = 0.2$  (light green). The lines represent means of twenty simulations.

probability of the genes in each gene set being randomly distributed around the threshold,  $p_0$ , was calculated.

$$p(X=x) = \frac{\binom{n}{x} \binom{m}{N-x}}{\binom{n+m}{N}},$$

where  $x$  is the number of affected genes in the gene set,  $N$  is the total number of affected genes,  $n$  is the total number of genes in the gene set and  $m$  is the total number of genes not in the gene set. In cases of several  $p_0$ , we chose  $p_0$  corresponding to the major peak of the curvature plot, or if two major peaks are close to each other, the one associated with the largest  $p_0$ .

The significant gene sets and the affected genes in the four comparisons are shown in supplementary tables 1–4.

When comparing *BRCA2* and sporadic, only three gene sets were significantly ( $p < 0.01$ ) altered, one consisting of genes involved in DNA repair. For *BRCA1*, 15 genes sets are significantly ( $p < 0.01$ ) altered. The genes involve DNA binding, phosphorylation and cell cycle. Among the six enriched gene sets in the comparison between familial (*BRCA1* and *BRCA2*) and sporadic tumors, the most significant enrichment is that of genes associated with the mitochondria. Almost as significant is the lack of electron transport genes among the selected genes.

We then tried to select genes using alternative methods. When controlling the FWER or the Benjamini-Yekutieli FDR, no genes were selected at the level  $\alpha = 0.05$  and when using the Benjamini-Hochberg FDR only the comparison between *BRCA1* and *BRCA2* yielded a non-zero list of genes. At the control-level  $\alpha = 0.2$ , all four comparisons resulted in lists comprised of 2–450 genes, which is much less (familial versus sporadic) or comparable

(BRCA1 versus BRCA2) to the number of genes selected as having  $p < p_0$ . The significant genes sets after gene set analysis are partially overlapping with the results from the  $p_0$  threshold (supplementary tables 5–8).

For visualization of the distributions, we used a modified version of 'gene set enrichment analysis' [8], where the genes were sorted on the  $p$ -values to form a sequence. Genes belonging to the tested gene set are attributed a

value of  $\sqrt{\frac{m}{n}}$  and the remaining genes are given the value

$-\sqrt{\frac{n}{m}}$ . The cumulative sum of the sequence is calculated

and plotted for a few gene sets in figure 4. The more a gene set deviates from a uniform distribution, the greater absolute value of the sum. For these gene sets, the threshold at  $p_0$  coincides with the maximum of the cumulative sum.

## Discussion

High-throughput techniques such as gene expression profiling by microarrays allows rapid screening of large amounts of data simultaneously. However, the staggering amount of data produced causes new problems, such as how to determine a meaningful  $p$ -value threshold between differentially expressed and unchanged genes. Methods such as Benjamini & Hochberg's [1] or Benjamini & Yekutieli's [11] control false positives proportions, but yield no information about false negatives proportions. Our method can produce a full description of the probability density function for differentially expressed genes, which makes it possible to rationally choose a desired ratio between  $\pi_0$  and  $1 - \pi_0$  for which the  $p$ -value threshold can be calculated.

The most important features of our proposed  $p$ -value threshold are that it is data dependent and unsupervised. It relies on an independent estimate of the unchanged genes proportion,  $\hat{\pi}_0$ . The method works best for  $\pi_0 \in (0.20, 0.95)$  and  $\lambda > 5$ , situations in which the number of alternative features is non-negligible and has a minimum separation from the null features.

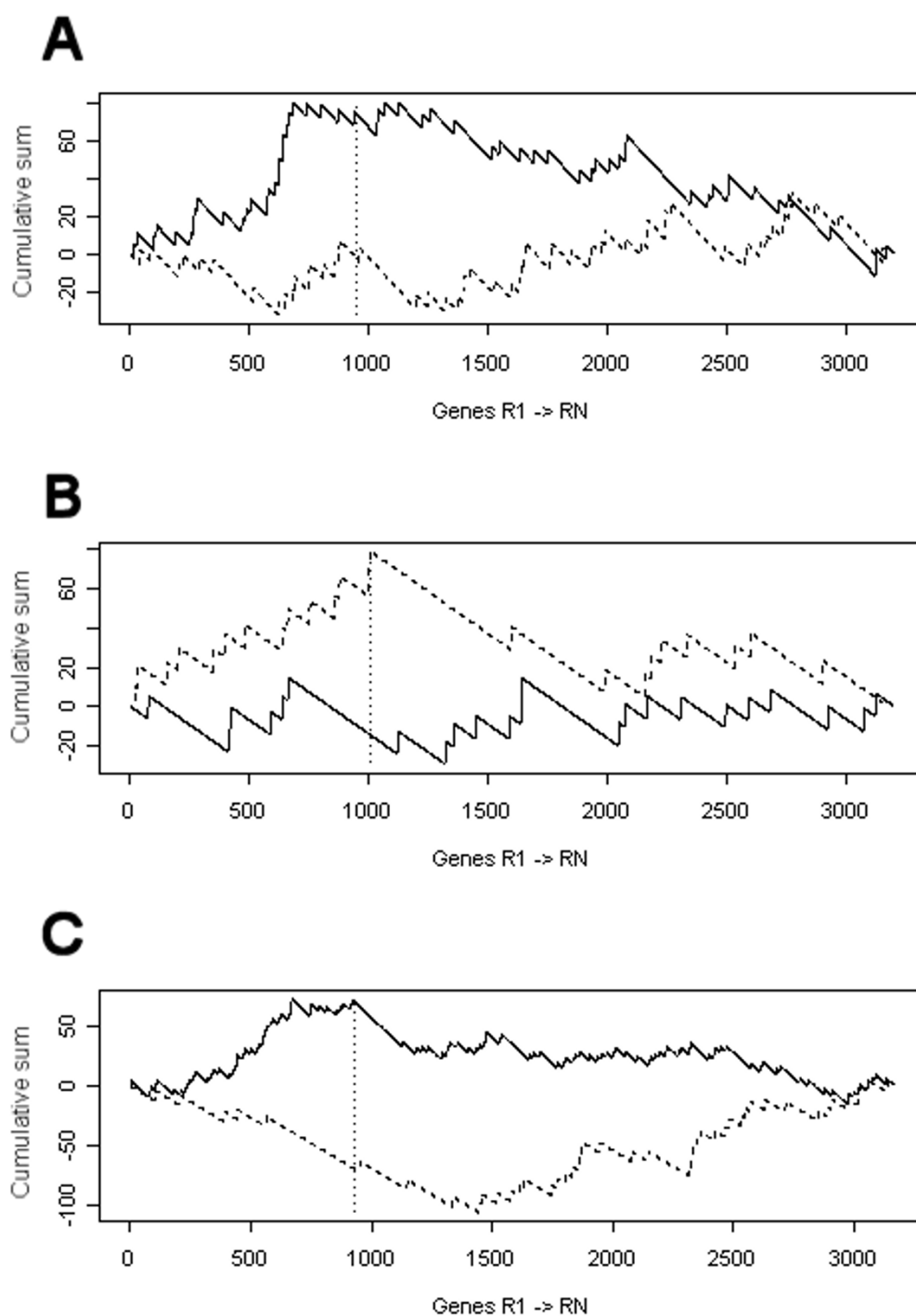
The value of  $\lambda$  will increase with the separation of the populations of alternative and null genes. The power of the statistical test will also be reflected on  $\lambda$ , as well as the quality of the data. On commercial platforms with multiple synthetic probes,  $\lambda$  tends to be higher than on microarrays spotted with PCR products (data not shown).

As the tolerable amount of false calls depends on the research question, relaxed selection criteria are needed for questions which require a balance between the inclusion of false positives and the exclusion of false negatives. Con-

ventional ways of selecting genes, such as stringent control of the FWER or FDR, select few, if any, false positives whereas the proportion of false negatives is at its maximum even with good separation between alternative and null features (figure 3). Only when relaxing the thresholds are we able to bring the two rates into equilibrium. Ideally, with no overlapping populations between truly alternative and truly null genes and a single population of alternative genes, the threshold at  $p_0$  will select  $1 - \hat{\pi}_0$  of the total genes as differentially expressed. However, in practice there is always some overlap in  $p$ -values between truly alternative and truly null genes. Selection of genes with  $p < p_0$  will frequently include many genes which can be useful under some circumstances. For instance, although in the examples of the case study the FDRs are high at  $p_0$ , almost all alternative genes are included, as can

be estimated by  $1 - \int_{p_0}^1 f_1(p)$ . In the functional analysis,

the plots of the cumulative sums show that setting the threshold at  $p_0$  is biologically valid. Up until these points there is still a contribution of genes in the relevant gene sets. A more stringent selection would neglect the evidence of many genes being marginally differentially expressed. By selecting genes through controlling the FDR, similar results can be obtained with relaxed thresholds. The analysis of biological function that we present on the BRCA data set was expected to reveal groups of genes related to the cellular processes that are affected by either mutant BRCA1 or mutant BRCA2. Indeed for certain predefined gene sets, e.g. genes involved in DNA repair, cell cycle progression or nucleic acid interactions, the proportion of genes with  $p < p_0$  is higher than can be expected from a random distribution. The observation that the threshold  $p_0$  coincides with the maximum of deviation from zero in figure 4 suggests that  $p_0$  is at the breakpoint where the enrichment of the gene set culminates. We noticed that there are more gene sets significantly affected by mutated BRCA1 than mutated BRCA2. Next to "protein phosphorylation", specifically gene sets involved in cell cycle regulation are enriched, including "cell cycle", and the somewhat less significantly induced gene sets "mitosis" and "spindle". As for mutated BRCA2, there is no indication that its impairment affects cell cycle progression; the genes in the gene set "mitosis" appear even to be depleted among the affected genes. The involvement of DNA repair is additionally endorsed by the less significantly induced gene sets "response to stress" and "induction of apoptosis by extracellular signals". From the literature we know that the BRCA network deals with lesions that block or interfere with DNA replication. BRCA1 has a role in DNA damage sensing but its precise function is not known (for review see [12]). BRCA2 is directly involved in DNA repair [13]. Interestingly, when BRCA1 and BRCA2 mutation positive tumors are com-

**Figure 4**

**Enrichment of gene sets in the BRCA-study.** Genes  $R_1, \dots, R_N$  ( $N > 3,000$ ) are ordered on their  $p$ -values and the cumulative sum is calculated to determine whether the members of a gene set are enriched. Starting with the top-ranking gene, the sum increases when a gene in the gene set is encountered and decreases otherwise. The gene sets shown here are **(A)** cell cycle (solid line – BRCA1-sporadic; dashed line – BRCA2-sporadic) and **(B)** DNA repair (solid line – BRCA1-sporadic; dashed line – BRCA2-sporadic). **(C)** For the comparison BRCA1/2-sporadic, the gene sets consisting of genes functional in the mitochondria (solid line) and electron transport (dashed line) are depicted. The vertical dotted lines correspond to the thresholds,  $p_0$ , for BRCA1, BRCA2 and BRCA1/2 compared to sporadic.



pared together against sporadic tumors, mitochondrial genes are affected but not electron transport genes. Until recently, the main function of mitochondria was thought to be the provision of energy for the cell through the creation of an electrical potential across its membrane. However, mitochondria are also involved in one route of apoptosis and recently there has been evidence for dysfunctional mitochondria being associated with premature ageing [14]. These are functions that could possibly be connected to the BRCA network.

## Conclusion

The method presented here will allow researchers to set unsupervised thresholds to select alternative features from mixed populations of  $p$ -values. As the discipline of systems biology evolves, there will be a need to compare global measurements of different levels (RNA, proteins, metabolites, etc.). The evidence of features being alternative can be used as weights in a comparison. In this context, a mathematical description of significant features enables a probabilistic approach to identify affected pathways.

## Method

### Setting thresholds

The  $p$ -values are defined as independent continuous random variables  $P_1, P_2, \dots, P_n$  taking values in the interval  $[0, 1]$ . Let us represent by  $f$  and  $F$  the probability density and cumulative probability functions of the distribution of a generic  $p$ -value  $P$ , respectively. We require that  $f$  is twice differentiable.

We represent by  $f_0(p)$  and  $f_1(p)$  the  $p$ -value densities for a null feature and an alternative feature, respectively. For a generic feature, its  $p$ -value density can be written as the mixture

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p), \quad \forall p \in [0, 1], \quad (1)$$

where  $\pi_0$  represents the proportion of null features out of the total under study. Note that  $f_0(p)$  takes the value 1 for all  $p$  within  $[0, 1]$ .

A common way of visualizing all  $p$ -values is to make a graph of the sorted  $p$ -values according to the features (see for example the solid line in figures 1B and 1D). If there are no alternative features, this line should roughly be a straight line with a 45 degrees angle. The presence of alternative features makes the line more convex, as seen in our examples. This line corresponds to  $F^{-1}$ , where  $F$  is the cumulative probability function as defined above.

We represent by  $p_0$  the  $p$ -value threshold defining the largest  $p$ -value corresponding to features identified as alternative. We shall define  $p_0$  as the maximum of the second

derivative of  $F^{-1}$ . The second derivative correspond to the curvature of the original function. Using (1), we get that  $p_0$  satisfies

$$f(p_0) \frac{d^2}{dp^2} f(p_0) = 3 \frac{d}{dp} f(p_0)^2 \quad (2)$$

Note that there need not be a unique value of  $p_0$  satisfying (2). The derivation of the equation is given in the appendix.

For an intuitive understanding, consider the extreme case where the alternative features under study all have  $p$ -values equal to zero, whereas the null features have a uniform distribution over  $[0, 1]$ . Then  $F$  has probability mass at  $p = 0$  equal to the proportion of alternative features ( $1 - \pi_0$ ), and in the interval  $(0, 1]$  is described by a straight line between  $(0, 1 - \pi_0)$  and  $(1, 1)$ . It is clear that the best threshold  $p_0$  is the one that selects all features with zero  $p$ -values. This corresponds to the turning point of  $F^{-1}$ . In practice, the alternative features will not all have zero  $p$ -values, but the turning point of  $F^{-1}$ , where its second derivative is zero, remains the best threshold to yield a compromise between false positives and false negatives proportions.

### Application to exponential distributions

To give some insight about how the proposed threshold works, let us consider the case when the alternative  $p$ -values have an exponential distribution. This consists of one of the simplest functional forms for a monotonically decreasing density within  $[0, 1]$  with probability density function given by

$$f_1(p) = G(\lambda) e^{-\lambda p}, \quad 0 \leq p \leq 1, \quad (3)$$

where  $G(\lambda) = \frac{\lambda}{(1 - e^{-\lambda})}$ ,  $\lambda > 0$ , a constant guaranteeing

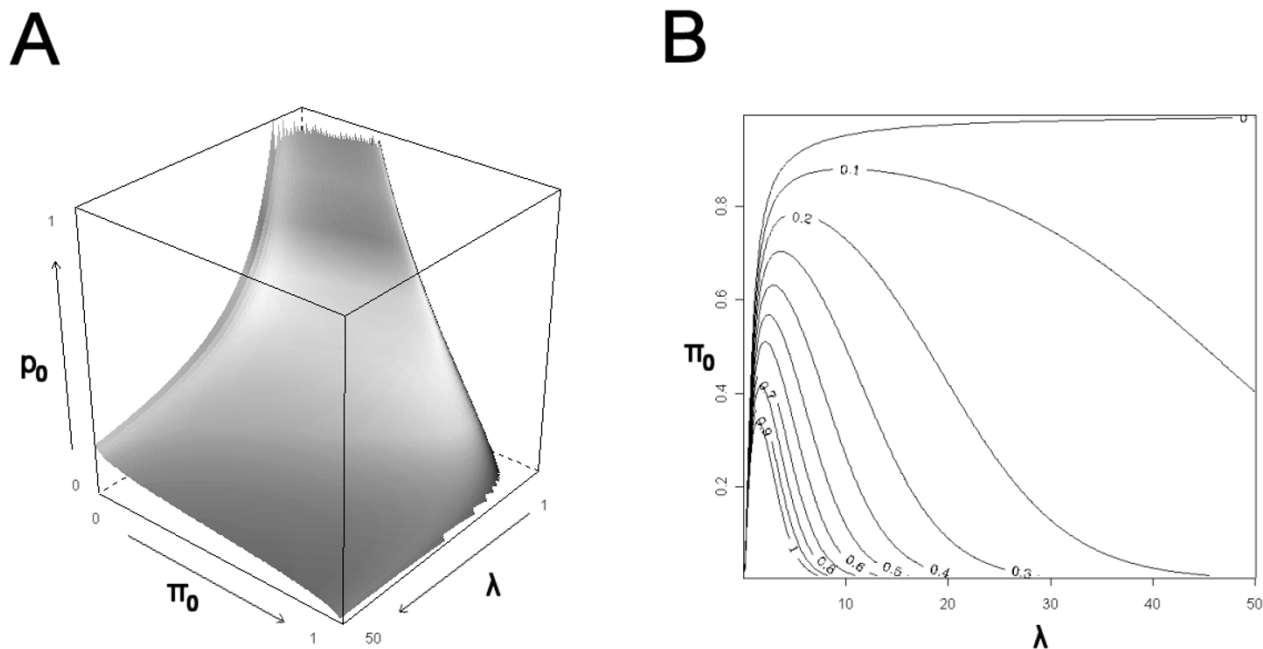
that  $\int_0^1 f_1(p) dp = 1$ . As  $\lambda \rightarrow 0$ ,  $f_1$  approaches the uniform distribution. Equation (2) yields

$$p_0 = \frac{\log(2G(\lambda)(1 - \pi_0)) - \log \pi_0}{\lambda} \quad (4)$$

If we replace this into (3), we obtain

$$f_1(p_0) = \frac{\pi_0}{2(1 - \pi_0)} \quad (5)$$

and, by replacing the latter into expression (1) for  $f$ , we get

**Figure 5**

**The  $p$ -value threshold dependencies on  $\pi_0$  and  $\lambda$ .**  $p$ -values come from a mixture of  $\pi_0$  null genes and a single population of  $1 - \pi_0$  alternative genes following an exponential distribution  $f_1(p) = \frac{\lambda}{(1 - e^{-\lambda})} e^{-\lambda p_0}$  visualized in (A) a 3D plot and (B) a contour plot of  $p_0$ .

$$f(p_0) = \frac{3\pi_0}{2} \quad (6)$$

Thus, the threshold is set at a point where the proportion of the alternative features is half the proportion of the null features. In other words, the genes with a  $p$ -value less than this threshold all have some evidence against the null hypothesis, whereas very few truly differentially expressed genes are excluded. For many purposes this definition gives a too wide selection, and it is more desirable to have a point where the proportion of alternative genes is higher than the proportion of null genes, e.g. at a point  $p_1$  where

$$f_1(p_1) = \frac{2\pi_0}{(1 - \pi_0)}.$$

At  $p_1$ , the proportion of alternative genes is twice the proportion of null genes. However, with the functional description given above, any point depending on the ratio between the proportions of null and alternative genes can be found in an unsupervised manner.

The dependency of the threshold  $p_0$  on  $\lambda$  and  $\pi_0$ , is visualized in figure 5. In this figure, it is clear that the suggested threshold  $p_0$  varies with  $\pi_0$  and  $\lambda$  in the desired way: as  $\lambda$  increases, the separation between the null and alternative  $p$ -values distributions increases and, therefore,  $p_0$  decreases. On the other hand, as the proportion of null genes  $\pi_0$  increases, naturally it becomes harder to identify the alternative genes, and again  $p_0$  decreases.

For practical reasons it is desirable to have an estimate of the proportion of false positives in the selection. Provided the described functions, we can integrate  $f_0$  and  $f$  in the interval  $[0, \max(p)]$ . The ratio between the two will be an estimate of the proportion of false positives.

Extensions to more than one alternative component are straightforward and given in appendix. Application of the more flexible beta distribution yield comparable results (see appendix).

### Non-parametric approach

In practice, there is interest in making as few distributional assumptions as possible so as to yield a robust approach. We suggest estimating  $F$  non-parametrically. This can be done by plotting the sorted  $p$ -values of all genes against their percentile rank (as the scatterplots in figures 1B and 1D) and estimating  $F^{-1}$  by fitting a smooth function – such as a cubic spline – to the data points. The smoothing parameters are chosen by cross-validation.

In non-parametric regression, the equivalent degrees of freedom for noise ( $EDF$ ) is defined by  $EDF = \text{tr}\{I - A(\alpha)\}$ , where  $A(\alpha)$  is the hat matrix associated with spline smoothing with smoothing parameter  $\alpha$ . The relation between  $EDF$ , the generalized cross validation ( $GCV$ ) score and the residual sum of squares [15,16] can be written as

$$\text{residual sum of squares} = \frac{GCV(\alpha) \times (EDF)^2}{n}. \quad (7)$$

The optimal  $EDF$ , and thus smoothing parameter, to fit the smoothing spline is found by minimizing the residual sum of squares.

To get the curvature of the spline,  $\hat{F}^{-1}$ , it is differentiated numerically twice. The critical points  $p_0$ , of this second derivative can be identified.

The R code and additional information is available at <http://www.medgencentre.nl/pla>.

### Authors' contributions

JPS designed and implemented the method and drafted most of the manuscript. RXM accounted for statistical expertise and substantial parts of the draft. MGG, IT and HV supervised the study and participated in manuscript writing. All authors read and approved the final manuscript.

### Additional material

#### Additional File 1

**Supplementary tables 1–4.** Gene sets enriched in the comparisons among familial (BRCA1 and BRCA2 mutation positive) and sporadic tumors selecting the genes with  $p < p_0$ .

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-6-177-S1.xls>]

#### Additional File 2

**Supplementary tables 5–8.** Gene sets enriched in the comparisons among familial (BRCA1 and BRCA2 mutation positive) and sporadic tumors, controlling the Benjamini-Hochberg FDR at  $\alpha = 0.2$ .

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-6-177-S2.xls>]

#### Additional File 3

**Supplementary figure 1.** Histograms and scatterplots of  $p$ -values from comparisons between BRCA2 mutation positive and sporadic tumors, and between BRCA1 and BRCA2 mutation positive and sporadic tumors.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-6-177-S3.bmp>]

#### Additional File 4

**Appendix.** Calculation of derivatives and the application to exponential and Beta distributions.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-6-177-S4.pdf>]

#### Additional File 5

**The pseudo-code and R code.** Methods to calculate  $p_0$  and estimate  $f(p)$  given a vector of  $p$ -values.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-6-177-S5.R>]

### Acknowledgements

The authors would like to thank Dr. P. Eilers at the Department of Medical Statistics, LUMC, The Netherlands, for helpful comments and critical reading of the manuscript.

### References

1. Benjamini Y, Hochberg Y: **Controlling the false discovery rate – a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57**:289-300.
2. Ferkingstad E, Langaas M, Lindqvist B: **Estimating the proportion of true null hypotheses, with application to DNA microarray Data.** *Preprint Statistics* 2003, **4**: [<http://www.math.ntnu.no/preprint/statistics/2003/>].
3. Grenander U: **On the theory of mortality measurement. Part II.** *Skandinavisk Aktuarietidskrift* 1956, **39**:125-153.
4. Schweder T, Spjøtvoll E: **Plots of  $p$ -values to evaluate many tests simultaneously.** *Biometrika* 1982, **69**:493-502.
5. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:9440-9445.
6. Begley TJ, Rosenbach AS, Ideker T, Samson LD: **Damage recovery pathways in *Saccharomyces cerevisiae* revealed by genomic phenotyping and interactome mapping.** *Molecular Cancer Research* 2002, **1**:103-112.
7. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344**:539-548.
8. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: **PGC-1 alpha-responsive genes involved in oxidative phosphorylation are**

- coordinately downregulated in human diabetes.** *Nature Genetics* 2003, **34**:267-273.
9. Segal E, Friedman N, Kaminski N, Regev A, Koller D: **From signatures to models: understanding cancer using microarrays.** *Nature Genetics* 2005, **37**:S38-45.
  10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**:25-29.
  11. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Annals of Statistics* 2001, **29**:1165-1188.
  12. Venkitaraman A: **Tracing the network connecting BRCA and Fanconi anaemia proteins.** *Nature Reviews Cancer* 2004, **4**:266-276.
  13. Yang HJ, Jeffrey PD, Miller J, Kinnucan E, Sun YT, Thoma NH, Zheng N, Chen PL, Lee WH, Pavletich NP: **BRCA2 function in DNA binding and recombination from a BRCA2-DSS1-ssDNA structure.** *Science* 2002, **297**:1837-1848.
  14. Trifunovic A, Wredenberg A, Falkenberg M, Spelbrink JN, Rovio AT, Bruder CE, Bohlooly Y, Gidlof S, Oldfors A, Wibom R, Tornell J, Jacobs HT, Larsson NG: **Premature ageing in mice expressing defective mitochondrial DNA polymerase.** *Nature* 2004, **429**:417-423.
  15. Green P, Silverman B: *Nonparametric regression and generalized linear models: a roughness penalty approach* London: Chapman and Hall; 1994.
  16. Hastie T, Tibshirani R, Friedman J: *The elements of statistical learning: data mining, inference, and prediction* New York: Springer; 2001.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

