entropy

MDPI

Article

Ordering Quantiles through Confidence Statements

Cassio P. de Campos [1,*,†], Carlos A. de B. Pereira [2,†], Paola M. V. Rancoita [3,†] and Adriano Polpo [4,†]

[1] Queen's University Belfast, Belfast BT7 1NN, Ireland, UK
[2] Institute of Mathematics and Statistics, University of São Paulo, São Paulo 05508-090, Brazil; cadebp@gmail.com
[3] Centre for Statistics in the Biomedical Sciences, Vita-Salute San Raffaele University, Milan 20132, Italy; rancoita.paolamaria@unisr.it
[4] Department of Statistics, Federal University of São Carlos, São Carlos 13565-905, Brazil; polpo@ufscar.br
[*] Correspondence: c.decampos@qub.ac.uk; Tel.: +44-(0)-289-097-6795
[†] These authors contributed equally to this work.

Academic Editor: Kevin H. Knuth
Received: 8 June 2016; Accepted: 26 September 2016; Published: 8 October 2016

Abstract: Ranking variables according to their relevance to predict an outcome is an important task in biomedicine. For instance, such ranking can be used for selecting a smaller number of genes for then applying other sophisticated experiments only on genes identified as important. A nonparametric method called Quor is designed to provide a confidence value for the order of arbitrary quantiles of different populations using independent samples. This confidence may provide insights about possible differences among groups and yields a ranking of importance for the variables. Computations are efficient and use exact distributions with no need for asymptotic considerations. Experiments with simulated data and with multiple real -omics data sets are performed, and they show advantages and disadvantages of the method. Quor has no assumptions but independence of samples, thus it might be a better option when assumptions of other methods cannot be asserted. The software is publicly available on CRAN.

Keywords: genomics; quantiles; median order; nonparametric; significance index

1. Introduction

A common analysis when dealing with -omics data sets of thousands of variables is to *rank* all these variables according to some measure that identifies how important they are to predict and/or retrospectively understand a certain disease characteristic or outcome. An example of a scenario is to study gene expression data of cancer patients and a class variable that identifies whether the patient relapsed or not. In this case, such analysis can be used as a first step of variable selection for the later application of other sophisticated statistical and/or biological experiments, as it may avoid expensive time-consuming analyses using uninteresting variables, or at least may help to prioritize them and save valuable resources [1]. Quor is a method designed to compare independent samples with respect to corresponding quantiles of their populations and to provide a confidence value for the order of such arbitrary quantiles. This confidence on the order of the quantiles can be used to build a ranking of importance for the variables in a domain. As an example, suppose we have two populations (healthy and ill subjects) and are interested in the median values of a variable representing the level of expression of a particular gene. The goal is to obtain the confidence that "the median expression of that gene in the first population is strictly smaller (respectively greater) than the median expression of the same gene in the second population". The comparison of medians might suggest that the gene is under- or over-expressed in the ill subjects, or simply that there is no significant difference of expression between the populations. By applying such computation to all genes in a data set, one can rank them based on the obtained confidence values.

Often, approaches for ranking variables may rely on unrealistic assumptions, such as normality of samples, asymptotic behavior of statistics, approximate computations, comparisons of only two populations/groups, need of equivalent number of samples in the groups (across distinct variables, not simply across groups), among others [2]. For instance, methods for hypothesis testing such as the Student's t-test and the Mann–Whitney–Wilcoxon's rank–sum u-test [3–5] have been widely employed for such purposes. Their statistics (or their p-values) are often used to sort variables into some ranking of importance. Arguably, they represent the most commonly used methods for this problem in biomedical applications (even if they were not originally devised for such purpose). Quor is nonparametric and assumes nothing but independence of samples, which may potentially increase its applicability and is particularly useful when assumptions of other methods cannot be asserted. It can deal with different numbers of samples and missing data, and yet can properly compare these variables. Its computations are very efficient using a simple dynamic programming idea and are performed using exact distributions with no need for any asymptotic consideration. Other ideas to work with quantiles do exist. For instance, [6] developed confidence intervals for different quantiles of a single population, while Quor targets two or more populations and only one quantile (abeit not necessarily the same) in each population.

Experiments with simulated data suggest that Quor might be a better option when the underlying assumptions of other methods are suspected not to hold. In particular, Quor achieves greater area under the receiver operating characteristic (ROC) curves than many competitors, such as t-test, u-test, Mood's median test, and others. We also employ multiple real benchmark -omics data sets to empirically compare Quor and other methods with respect to the ranking of the variables they provide. It is shown that the rankings produced by Quor have better relation to the accuracy of group prediction based on a univariate model (with the variable in question). Finally, a study with copy number data of diffuse large B-cell lymphoma patients demonstrates the use of Quor with quantiles other than the median and indicates that Quor may produce better ranking than other methods (including the standard method for that problem, that is, the Fisher exact test applied after the categorization of the variables).

2. Methods

We describe here the details of Quor and present an efficient algorithm for its computation. The method is built on the ideas of confidence statements developed long ago [7,8] and revisited more recently [9]. The proposed method uses nonparametric confidence intervals for quantiles based on the binomial distribution [10]. Its goal is to compute a confidence value indicating how much one believes that quantile parameters of different populations/groups are ordered among themselves. This confidence can be loosely related to standard confidence regions used in frequentist analysis, as we explain later on. We do not assume any particular quantile nor a specific number of populations, even though the case of comparing medians of two populations is arguably the most common scenario for its application. For ease of expose, we will present the method using two groups, but the extension to three or more groups is straightforward.

The problem is defined as follows. Let Q_1 and Q_2 represent, respectively, the quantiles at arbitrary percentages q_1 and q_2 for two populations, that is,

$$\Pr(\{X_j \leq Q_j\} \mid Q_j) \geq q_j \text{ and } \Pr(\{X_j \geq Q_j\} \mid Q_j) \geq 1 - q_j, \tag{1}$$

with $j = 1, 2$ (these inequalities are tight in the continuous case). Let $\mathbf{x}_j = (x_j^{(1)}, \ldots, x_j^{(n_j)})$ be a sorted sample of size n_j from population $j = 1, 2$. The goal is to compute a confidence value in $[0, 1]$ that indicates how much we believe in the statement $Q_1 < Q_2$ (or similarly $Q_2 < Q_1$). This value can and will be used later to rank variables in an order of the confidence that the underlying populations have such difference in their quantiles.

Let $(X_j^{(1)}, X_j^{(2)}, \dots, X_j^{(n_j)})$ be the sorted vector of n_j independent and identically distributed random variables from population j (random variables from distinct populations are not necessarily identically distributed). Since the probability of one observation being smaller than the population quantile Q_j is q_j, it is straightforward that it holds for $X_j^{(i)}$ (that is, the ith ($i = 1, 2, \dots, n_j$) order statistics):

$$\Pr(\{X_j^{(i)} \leq Q_j\} \mid Q_j) \geq \sum_{k=i}^{n_j} \binom{n_j}{k} q_j^k (1 - q_j)^{n_j - k}, \tag{2}$$

and

$$\Pr(\{X_j^{(i)} \geq Q_j\} \mid Q_j) \geq \sum_{k=0}^{i-1} \binom{n_j}{k} q_j^k (1 - q_j)^{n_j - k}. \tag{3}$$

These inequalities come from probabilities obtained with a binomial distribution with n_j trials and probability of success q_j (they are inequalities because of possible ties at Q_j in the discrete case only). Consider a pair of order statistics $(X_1^{(i_1)}, X_2^{(i_2)})$ and the event $\mathrm{E} = \{Q_1 \leq X_1^{(i_1)}\} \wedge \{X_2^{(i_2)} \leq Q_2\}$. Given the independence assumptions, one can compute

$$\Pr(\mathrm{E} \mid Q_1, Q_2) = \Pr(\{X_1^{(i_1)} \geq Q_1\} \mid Q_1) \cdot \Pr(\{X_2^{(i_2)} \leq Q_2\} \mid Q_2) \tag{4}$$

using the product of binomial probabilities from Equations (2) and (3) (in the discrete case, a lower bound for the probability is obtained).

The samples $\mathbf{x}_1$ and $\mathbf{x}_2$ are assumed to be sorted non-decreasingly (one could easily sort them beforehand if they were not). After these samples are observed, the only unknown quantities of interest are the quantiles Q_1 and Q_2 of the two populations being studied. By replacing random variables with their observations, we create the statement $\mathrm{e}(i_1, i_2)$ as follows:

$$\mathrm{e}(i_1, i_2) = \{Q_1 \leq x_1^{(i_1)}\} \wedge \{x_2^{(i_2)} \leq Q_2\}, \tag{5}$$

which has confidence given by Equation (4). Note that the statement e is only a function of the order statistics i_1 and i_2 and not of the actual observed values of $X_1^{(i_1)}$ and $X_2^{(i_2)}$, and these order statistics are only what is needed in order to compute using Equation (4) (assuming that the data are available at hand). At this point after sampling, the value of Equation (4) becomes a confidence value instead of a probability [7]. This confidence regards the unknown quantities of interest, in our case the parameters Q_1 and Q_2. If we take each part of the statement e in Equation (5) separately (namely, $\{Q_1 \leq x_1^{(i_1)}\}$ and $\{x_2^{(i_2)} \leq Q_2\}$), then the corresponding values obtained for these statements, computed through Equations (2) and (3), are actual confidence values in the frequentist statistics jargon. Because of the independence assumption between the samples, we take their product as the confidence of the statement e itself.

Now, the idea is to look for statements $\mathrm{e}(i_1, i_2)$ that are able to tell us something about the order between Q_1 and Q_2. With a quick inspection of Equation (5), we have

$$\{x_1^{(i_1)} < x_2^{(i_2)}\} \wedge \mathrm{e}(i_1, i_2) \Rightarrow \{Q_1 < Q_2\}, \tag{6}$$

that is, the assertion in the left-hand side of Equation (6) implies an order for the quantiles, thus its confidence is a lower bound for the confidence of the right-hand side. Because we know how to compute the confidence value of $\mathrm{e}(i_1, i_2)$ through Equation (4), and because any time the assertion $x_1^{(i_1)} < x_2^{(i_2)}$ is false the Equation (6) is not applicable, we run over all pairs (i_1, i_2) of orders such that

$x_1^{(i_1)} < x_2^{(i_2)}$ and keep the maximum confidence value of the e statements built from these pairs as our estimation for the confidence that $Q_1 < Q_2$ holds true:

$$\max_{i_1,i_2:\, x_1^{(i_1)} < x_2^{(i_2)}} \mathrm{Conf}(\mathrm{e}(i_1, i_2)) \le \mathrm{Conf}\left(\{Q_1 < Q_2\}\right). \tag{7}$$

Hence, we use the left-hand side of Equation (7) as our confidence about $Q_1 < Q_2$. We can take the maximum value because each confidence value $\mathrm{Conf}(\mathrm{e}(i_1, i_2))$ is an actual confidence for $\mathrm{Conf}\left(\{Q_1 < Q_2\}\right)$; thus, it is possible to use the one yielding the greatest confidence. The value, however, might only be an approximation to $\mathrm{Conf}\left(\{Q_1 < Q_2\}\right)$ because of the gap between $x_1^{(i_1)}$ and $x_2^{(i_2)}$. Since we want to obtain a maximal value, a mathematical property that speeds up computations is to choose i_2 such that $x_2^{(i_2)}$ is the smallest possible value greater than $x_1^{(i_1)}$, that is, the value of i_2, in order to maximize the confidence, is uniquely and easily computable from the value i_1 (this holds because there is no reason to leave a larger gap between $x_1^{(i_1)}$ and $x_2^{(i_2)}$ if a smaller gap is possible, as smaller gaps will certainly yield higher confidence values). Hence, we only need to optimize over the possible values of i_1 to find the highest confidence value. Such a procedure is presented in Algorithm 1. We recall that if one wants to compute the confidence of $Q_2 < Q_1$, then they simply need to rename the variables accordingly before invoking the algorithm, while if one wants to check for every possible order of the quantiles (for instance, when we are not interested in a particular order a priori), then both permutations are to be checked.

Algorithm 1 (Quor Core) Confidence value of a statement about the ordering of quantile parameters of two populations.

```
Input  a data set with samples $\mathbf{x}_j = (x_j^{(1)}, \ldots, x_j^{(n_j)})$, for $j = 1, 2$, and the quantiles of interest $q_1$ and $q_2$.
Output  the confidence value that the statement $Q_1 < Q_2$ holds true.

1 Pre-compute the values that appear in Equations (2) and (3) by making a cache. For $j = 1, 2$:
  $\mathrm{cache}(j, 0) \leftarrow (1 - q_j)^{n_j}$ and for $i = 1, \ldots, n_j$:
      $\mathrm{cache}(j, i) \leftarrow \mathrm{cache}(j, i-1) + \binom{n_j}{i} q_j^i (1 - q_j)^{n_j - i}.$
2 Return
      $\max_{1 \le i_1 \le n_1} \left(\mathrm{cache}(1, i_1) - (1 - \mathrm{cache}(2, \mathrm{argmin}_{1 \le i_2 \le n_2}(x_1^{(i_1)} < x_2^{(i_2)})))\right).$
```

Theorem 1. *Algorithm 1 uses space and time $O(n)$, with $n = n_1 + n_2$.*

Proof. Step 1 pre-computes the partial binomial sums. By doing it in a proper order, this can be accomplished in constant time for each $\mathrm{cache}(i, j)$, and the loop will execute $O(n)$ times. Step 2 performs a very simple dynamic programming, with $O(n_1)$ steps in the outer loop. The inner argmin can be accomplished with an overall total of $O(n_2)$ steps by keeping a pointer to the last found i_2 of each outer loop (because the vectors are sorted, the next result of argmin cannot be smaller than that, and so only one pass over each possible i_2 is done). Hence, Step 2 takes time $O(n)$ in the worst case too. □

Algorithm 1 is very fast (as fast as hypothesis testing methods such as the Wilcoxon rank–sum test). The correctness of Algorithm 1 comes from its simple dynamic programming formulation. At each loop of Step 2, we find a pair (i_1, i_2) yielding the highest possible confidence value for that given i_1, and the loop iterates over all possible i_1. It is worth noting that some computations in Algorithm 1 could suffer from numerical problems. We have addressed this issue by implementing it using incomplete beta functions and/or arbitrary precision in particular cases where n is quite large (many hundreds). Because of caching, this does not slow down the whole approach in a perceivable way.

The confidence value obtained with Quor provides information about differences in quantiles as well as similarities in quantiles (for example, in the case that confidence is low in both directions). This is in contrast to usual hypothesis testing, where no evidence in favor of the null hypothesis can be obtained. Finally, we point out that the ideas presented here can be extended to any number of groups. Derivations and algorithm implementation become more intricate, but overall computational complexity is still low. In order to give a sketch of the version with any number of groups, the dynamic programming needs to keep track of the order statistics of each of the groups in Step 2, which needs to further iterate over the m groups, achieving worst case time complexity bounded by m times the complexity for two groups. The complexity does not considerably increase (becomes only quadratic in m) because the dynamic programming keeps track of all optimal candidates for previously processed groups (there are at most m of them). We omit such more sophisticated versions for ease of exposition, and for the fact that working with two groups is, by far, the most common and interesting case, as we discuss in the sequel.

3. Results

The experiments are divided in simulated and real data studies with multiple -omics data sets. We use the first part of this section to discuss the benefits of Quor in terms of prediction accuracy with simulated data. Then, we use multiple data sets to show the quality of the produced ranking of Quor and widely used methods.

3.1. Simulation Study

We compare Quor to a number of competitors, namely (i) the two-sample t-test, which performs an analysis of means by comparing their normalized difference to a t distribution; (ii) the Mann–Whitney–Wilcoxon u-test, which checks whether the probability of an observation from one distribution being greater than an observation from another is equal to 1/2; (iii) the modified t-test as presented in the package `samr` [11]; (iv) the standard Mood's median test; and (v) the q-test for quantile testing [12]. Ranking covariates by the statistics of such tests is reasonable as long as the number of data samples is constant over all covariates (in which case the ranking can be equivalently done by sorting p-values). Each test has a different assumption about the distributions under null/alternative hypotheses. We choose as target quantiles for Quor the medians (so we are interested in ordering the medians of the populations) unless stated otherwise.

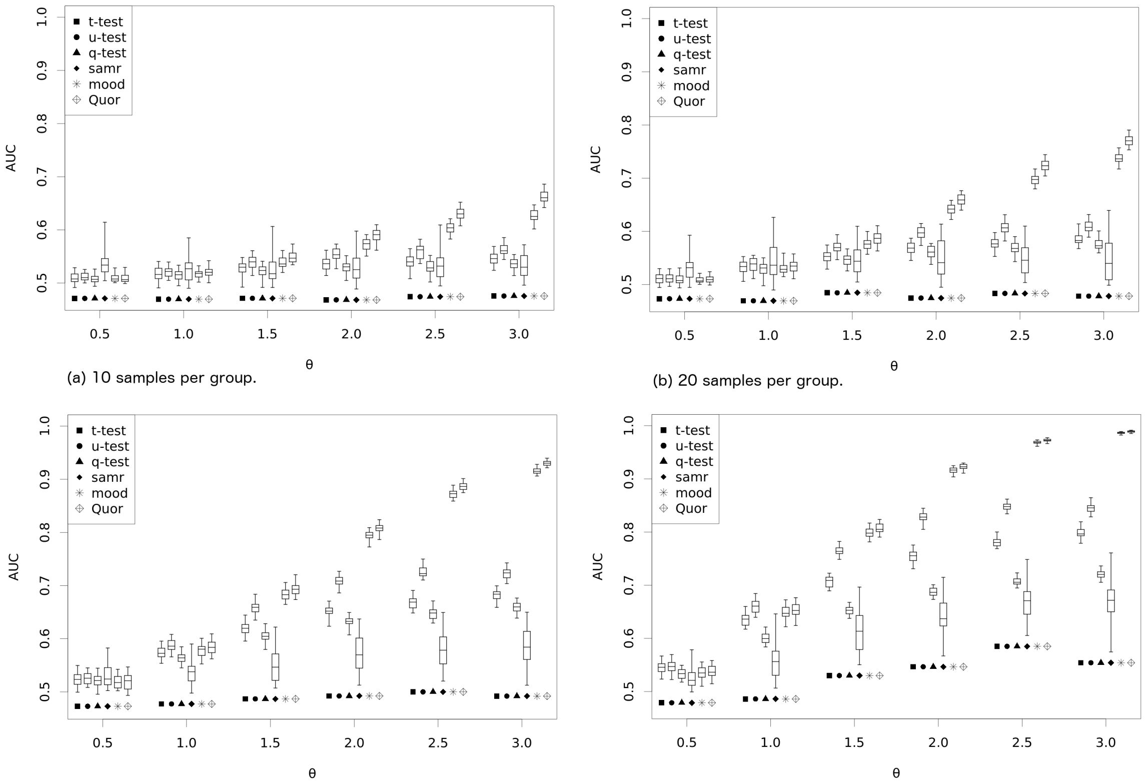
The goal of these experiments is to understand the goodness of the confidence value of Quor to identify the difference in the populations. We will draw a parallel to the p-values obtained by t- and u-tests. It is obvious that these procedures should only be used in cases where their assumptions hold. However, verifying whether their assumptions hold is by itself subject to issues. Bear in mind that we will perform analyses without trying to check if their assumptions are satisfied, and, in fact, we will analyze their behavior exactly in those cases where they are not.

To make results comparable, we study the area under the receiver operating characteristic (ROC) curve, as it is a measure (independent of any chosen threshold) for the prediction accuracy of the methods in identifying whether a gene is differently distributed in the two given groups. As argued recently by many statisticians [13], this approach for comparison of methods does not force the use of a particular significance level without justification. In the case of Quor, a high confidence value (above a certain threshold) would conclude that there is a difference (in that gene) between populations, while for t- and u-tests, a small value of the p-value (below a certain threshold) is used to achieve such a conclusion. The experiments consist of generating samples either from the same distribution for the two groups (in this case, we hope that the methods will not identify any difference) and from two different distributions (in which case, the methods should identify the difference). The ROC curve is built by varying the threshold value that each method uses to decide whether the samples are from the same or different distributions. In order to generate different distributions for the groups, we use the following two scenarios: an experimental setting with a mixture of Gaussians (a common assumption in biomedical analyses, including gene expression) and another with Gaussians. The amount of difference between the generating distributions of the two groups will be quantified by a single parameter θ, as explained in the continuation.

As a first experiment, all methods are run to identify the (nonexistent) difference between samples generated from a Gaussian with mean zero and standard deviation two (for each of the two groups). This is repeated 100 times with 2000 simulated genes each time (1000 in each group), and the results from the methods are recorded. The two groups are sampled as follows: (i) the first group from a mixture of two Gaussians with means equal to $\mu_{1,1} = -\theta$ and $\mu_{1,2} = \theta$ (both Gaussian standard deviations $\sigma = 2$) and with weights 2/3 and 1/3 for the mixture, respectively; and (ii) the second group from a mixture of two Gaussians with weights 1/3 and 2/3 for Gaussians with means $\mu_{2,1} = -2 \cdot \theta$ and $\mu_{2,2} = \theta$ (standard deviations again equal to two). All methods are run over the 2000 genes and results are recorded. Each of these 2000 points indicates whether a method successfully identified if the curves were the same (first thousand points) or different (next thousand points), so we have the amount of true/false positives and true/false negatives. Now, the area under the ROC curve is calculated using these 2000 values for each method. The high number of evaluations gives us a very precise picture of the area under the curve for each of the methods. This procedure is repeated 100 times and boxplots are constructed.

Our choice of mixture distributions is designed to produce faster variation in the medians than in the means when one increases the single experimental parameter θ. Hence, with the increase of the θ, the first group will have its mean and median decreasing, while the second will experience its mean stalled and its median increasing. Results are shown in Figure 1 over the variation of the θ from 0.5 to 3 that is used to define the generating distributions (for each value of θ, the area under the ROC curve is computed using 2000 points and repeated 100 times). The experiment is conducted for sample sizes (same for the two groups) in $\{10, 20, 50, 100\}$. From the figure, it is clear that Quor produces better results (larger area under the curve) than the other methods because the difference in the medians with the increase in θ grows faster than the difference in the means (so this is nothing but expected). Moreover, Quor does better than u-test (which should account better for the difference in medians than the t-test), q-test and Mood's median test, which are designed to check differences in medians.

While in the first batch of experiments we have used mixtures of Gaussian with medians that varied faster than means (thus benefiting Quor and the other quantile methods), in the second batch, we perform experiments with Gaussians so as to see whether t-test considerably dominates the other methods (as in this case its assumptions are fulfilled and the t-test is expected to perform better). Again, methods are run to identify the (nonexistent) difference between samples generated from a Gaussian with mean zero and standard deviation two (100 repetitions with 1000 genes each). We generate the groups from (i) a Gaussian with mean $\mu_1 = -\theta$; and (ii) a Gaussian with mean $\mu_2 = \theta$ (standard deviations always as $\sigma = 2$). One hundred repetitions are run once more with 2000 genes each, so again we have 1000 cases where no difference between groups should be identified and 1000 cases where a difference should be identified by the methods. Results of the area under the curve are displayed in Figure 2. In this case, and as expected, t-test and u-test perform better, with t-test slightly superior, although the difference is not so prominent. We argue that if one does not know whether the data are Gaussian, Quor might be a better choice. It produced better results when data were from a mixture of Gaussians, while not losing too much accuracy in the Gaussian experiment. Even though it produced worse results in this latter case, Quor directly returns which group has the greater quantile without any additional effort, which might be an extra benefit.



(a) 10 samples per group.

(b) 20 samples per group.

(c) 50 samples per group.

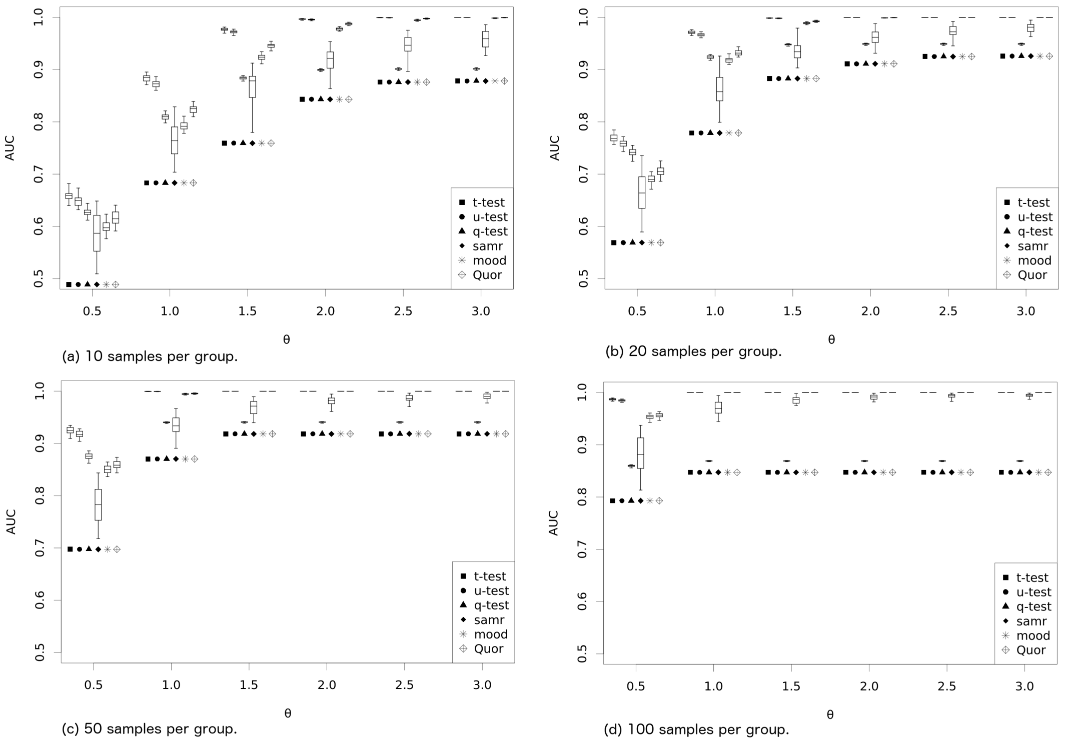
(d) 100 samples per group.

Figure 1. AUC comparison using mixture of Gaussians, area under the ROC curve for different methods. Samples of the two groups come from mixtures of two Gaussians (all standard deviations are $\sigma = 2$). The mixture of first group has Gaussians with means $\mu_{1,1} = -\theta$ and $\mu_{1,2} = \theta$ and mixture weights 2/3 and 1/3. The mixture of the second group has weights 1/3 and 2/3 with Gaussians with means $\mu_{2,1} = -2 \cdot \theta$ and $\mu_{2,2} = \theta$. Simulation is repeated 100 times with 2000 genes each, with number of samples as indicated in the subfigures. The control scenario (no difference in the groups) is built with both groups generated from a Gaussian $\mu = 0, \sigma = 2$.

3.2. Comparison and Discussion on Rank Evaluation in -Omics Data Sets

The analysis of -omics data deals with several thousands (or millions) of variables, which, for instance, may represent the expression of genes or the copy number of genomic regions. When the study is performed genome-wide, it is a common procedure during the analysis to somehow sort the variables with respect to their "performance", in order to select which of them should be considered in further investigations/experiments. This is especially important in recent data where the number of variables has become extremely large and sophisticated analyses over all variables would require prohibitive computational resources.

In this section, we apply the same methods as before except for the q-test (namely t-test, u-test, modified t-test from `samr`, Mood's median test and Quor) to the task of ranking the variables in some order of importance/preference using multiple -omics data sets. The q-test has been discarded from this and forthcoming experiments because of its poor quality in the simulated data and its terrible computational performance. The rank of the variables is defined for each method in the following way. When using t-, u- or Mood's median tests, we sort the variables by their p-values (in increasing order, as generally done in the literature), for the modified t-test of `samr` we use its absolute score, while, for Quor, we sort them by the confidence values (decreasingly).



(a) 10 samples per group.

(b) 20 samples per group.

(c) 50 samples per group.

(d) 100 samples per group.

Figure 2. AUC Comparison using Gaussians, area under the ROC curve for different methods. Samples of the two groups come from Gaussians with standard deviation $\sigma = 2$. The first group has a Gaussian with mean $\mu_1 = -\theta$ and the second group has $\mu_2 = \theta$. Simulation is repeated 100 times with 2000 genes each, with number of samples as indicated in the subfigures. The control scenario (no difference in the groups) is build with both groups generated from a Gaussian $\mu = 0, \sigma = 2$.

We consider two main applications. Firstly, we compare the quality of the obtained ranks by using eight benchmark data sets of gene expression profiling and proteomic spectra. In this situation, the rank of each variable is assessed on the basis of its capability in group classification [1]. Secondly, we apply the appropriate methods to a copy number data set of diffuse large B-cell lymphoma (DLBCL) patients, and we compare the ranks (yielded by different approaches) for well-established aberrations characterizing the three cell-of-origin subtypes.

3.2.1. Performance in Group Prediction

We compare the obtained ranking from the different methods on a collection of benchmark data sets. Table 1 presents the data sets used for this comparison. The data were obtained from internet repositories, and we defer further details to their corresponding citations (see Table 1).

For the purpose of evaluating the quality of the obtained ranking (yielded by each method), we would need to know the correct rank of the variables, which is unavailable in real data and cannot even be easily simulated (as the true ranking for a data set is a characteristic of the available samples and not of an underlying distribution). To overcome this drawback, we follow a previous study [1] and employ as our target the ranking obtained by sorting variables according to their accuracy when used alone for group prediction, that is, for each variable we compute the zero-one loss (prediction error) of a simple classification model that predicts which is the group of each individual using only that single variable. This is done using the whole data set and forms the ranking to which we will compare all other methods. We use the R package rpart, but any similar idea would suffice, as the information available to build the model leads to splitting the image of the predictive variable into pieces—we allow at most two splits to avoid an obvious overfit. We call this ranking the target ranking.

Table 1. Characteristics of eight benchmark data sets.

Data Set	$\mathbf{n_1 + n_2}$	# Var.	Type	Groups
Breast Cancer	46 + 51	24,481	GEP	Relapse/Not
Central Nervous System	21 + 39	7129	GEP	Dis. Subtype
Colon Cancer	22 + 40	2000	GEP	Disease/Not
Leukemia	47 + 25	7129	GEP	Dis. Subtype
Lung Cancer	24 + 15	2880	GEP	Relapse/Not
Ovarian Cancer	162 + 91	15,154	PS	Disease/Not
Prostate Cancer	8 + 13	12,600	GEP	Relapse/Not
Schizophrenia	34 + 32	20,992	GEP	Disease/Not

The sources of the data sets are as follows: Breast Cancer [14], Central Nervous System [15], Colon Cancer [16], Leukemia [17], Lung Cancer [18], Ovarian Cancer [19], Prostate Cancer [20], Schizophrenia [21]. $n_1 + n_2$ is the number of patients, shown by the amount in each group. PS stands for *Proteomic spectra*, GEP for Gene Expression Profiling.

We evaluate each method by taking the k best ranked variables according to it (for arbitrary different values k, as shown in the horizontal axis of Figure 3) and by computing the average prediction accuracy that such k best variables have according to the target ranking. A better method will result in lower values, since the prediction accuracy of the variables that are within the k best variables will reflect how the obtained ranking identifies the important variables which are exactly those with small rank value in the target ranking). For clarity of presentation, we divide such averaged accuracy value by that obtained by the target ranking, which is obviously optimal in this respect. The results in Figure 3 (for the data sets described in Table 1) show how far (in percentage) the named method is from the target rank. A value of zero in the curve means that the ranking was as good as the target rank, and higher values indicate worse performance, as the subset of variables has larger (averaged) prediction error (the correct ranking is assumed to have the optimal order in terms of prediction error). Each point in the graph is obtained by subsampling half of the data available for each gene and group. This is repeated 20 times, and the graph shows mean values and standard deviations (as bars). Hence, the methods have access only to part of the data, while the computation of the target ranking has been done with the whole data set.

Quor has different characteristics from the other methods, in particular from the t-test. Results in Figure 3 show an overall worse performance of the t-test (as well as of its modified version from `samr`), which has produced ranks with less prediction accuracy in almost every data set and almost every number of selected variables (suggesting that the produced rank is not very related to group prediction accuracy). Quor, Mood's median and u-test achieved better results, with Quor showing a clear superiority in three or four data sets and u-test in only one. Mood's median test show good results but is consistently outperformed by Quor.

3.2.2. Identifying Well-Known Lesions in Copy Number Data

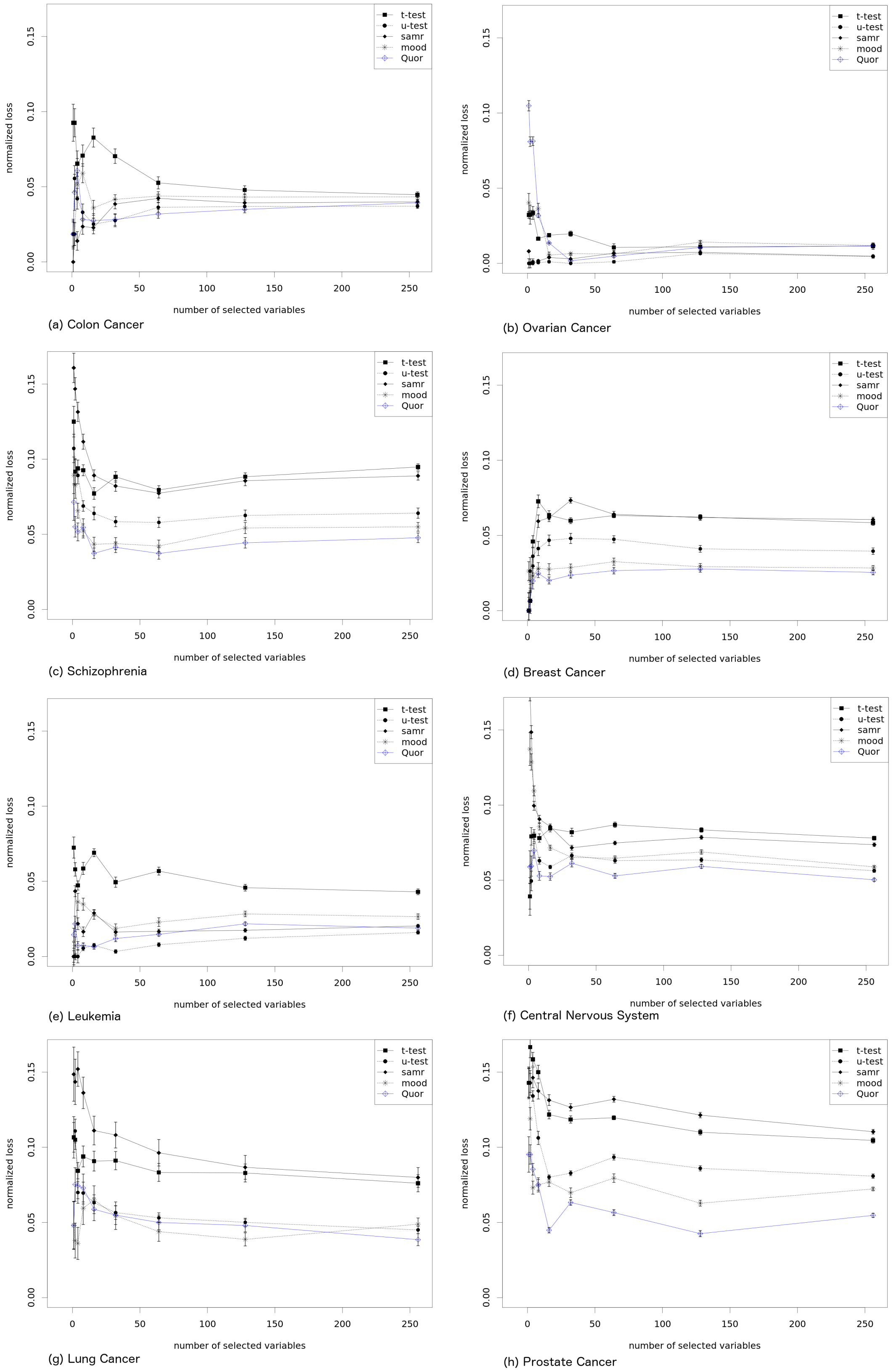
As a second application, we analyze the copy number (CN) microarray data of the 176 DLBCL patients of [22], for which the classification in the three following cell-of-origin subtypes was available: 71 germinal center B-cell like, or GCB; 74 activated B-cell like, or ABC; and 31 primary mediastinal B-cell lymphoma, or PMBL. DNA copy number aberrations are defined as genomic regions with a number of copies of DNA different from two (which is the normal value for autosomal chromosomes). Thus, the CN profile of a patient can be represented as a piecewise constant function, where segments showing an increased CN are called gains, while the ones with a decreased CN are called losses. The microarray CN data contain continuous values, due to both technical and biological reasons, and after applying a segmentation method, the regions of gains and losses are usually detected by using some thresholding (defined opportunely). As for this study, the data of [22] were obtained segmented and both with and without discretization in gains/losses/normal CN, from the Progenetix database [23]. Quor, t-test and u-test are applied to the continuous segmented data (with the goal of demonstrating the differences with respect to Quor), while the Fisher exact test (that can be referred as a standard method for the analysis of CN data) to the already discretized data. Mood's median test is not used as it has been consistently outperformed by Quor.

Heterogeneous diseases, like DLBCL, are characterized by aberrations that might be present only in 15%–35% of the cases, and disease subtypes may share some lesions [22,24]. Therefore, when testing a possible aberration related to DLBCL subtypes, we consider each possible test of two subtypes versus the remaining one, in order to find whether the aberration is prominently associated with any single one of the groups. Moreover, due to the low frequency of lesions in these data [22,24], Quor can be used in an opportune way by comparing quantiles of the distributions other than the median. For example, a gain may be identified associated with a group when the 85th quantile of its distribution is higher than the 95th quantile of the distribution of the other group (so we use $q_1 = 95\%$ and $q_2 = 85\%$ and compute the confidence of $Q_1 < Q_2$), while a loss may be identified when the 15th quantile of a group is lower than the 5th quantile of another (in this case, we use $q_1 = 15\%$ and $q_2 = 5\%$). Hence, Quor has been configured to identify differences in the tails of the distributions, which shows its ability to test different characteristics according to one's needs. We also include in the analysis the method Quor with median as target quantile for comparison. The Fisher exact test is performed as usual by testing a single type of aberration at a time (gain versus not gain and loss versus not loss). In the analysis, we consider all the genomic regions with at least five probes, obtained by the union of all segmented profiles. For each region and for each method, all evaluations are performed, and the best result is chosen to define the p-value or the confident value used in the ranking (multiple test correction is unnecessary, as orders are preserved).

Table 2. Well-known lesions associated with DLBCL cell-of-origin subtypes.

Gene	Cytoband	Association with Cell-of-Origin Subtypes [24]
REL	2p16.1	gain for GCB and PMBL
PRDM1	6q21	loss for ABC (it is not known for PMBL)
CDKN2A	9p21.3	loss for ABC
JAK2	9p24.1	gain for PMBL
BCL2	18q21.33	gain for ABC and in a minority of GCB and PMBL
SPIB	19q13.33	gain for ABC

Notice that in this setting the variables (i.e., the genomic regions) are highly dependent (for example, the first 20 regions identified by Quor when comparing the tail-related quantiles are all at cytobands 9p24 and 9p21). Nevertheless, the five approaches for ranking the regions are compared with respect to the yielding rank achieved in well-established CN lesions (shown in Table 2) associated with DLBCL cell-of-origin subtypes. The results are showed in Table 3, where ranks for each of those well-known lesions were identified. This study regards a common situation in the analysis of this type of data: regions are sorted according to some measure of importance and further analyses are performed with those regions which appear first (with low rank). With that in mind, Quor, using the tail-related quantiles, identified all considered lesions up to rank 1190 (with no contrast to established knowledge of the disease), while other methods needed a rank higher than two to four thousand. While Quor with tail-related quantiles performed well, Quor with medians was clearly less suited to the task, leading to high ranks and results in contrast with established knowledge.



(a) Colon Cancer

(b) Ovarian Cancer

(c) Schizophrenia

(d) Breast Cancer

(e) Leukemia

(f) Central Nervous System

(g) Lung Cancer

(h) Prostate Cancer

Figure 3. Ranking Quality. Average zero-one loss of the univariate classifiers obtained with each of the selected variables according to the ranking method (we always select them according to such rank). The values are shown with respect to the target ranking. Hence, zero means optimal and 0.15 means 15% worse than the target ranking, for instance. Values are means over 20 subsamples with half of the data for each gene and group. Bars show the standard deviation.

Table 3. Order statistics of well-known lesions associated with DLBCL cell-of-origin subtypes, obtained with several methods: Quor comparing either particular quantiles for the tails or medians, Fisher exact test, *t*-test and *u*-test.

Gene	Quor: Quantiles Comparing Tails	Quor with Medians	Fisher Exact Test	Student's *t*-test	Wilcoxon *u*-test
REL	1132	4645	2070	468	3903
PRDM1	435	1776	2294	24	986
CDKN2A	112	4130 (a)	302	1044	3565
JAK2	174	1345	127	370	884
BCL2	640	3678	- (b)	574	3655
SPIB	1190	237 (a)	232	3272	1065 (a)

(a) The result of the method is in contrast to the established knowledge about that lesion; (b) The test is not performed, because all patients have normal copy number in the discretized version of the data.

4. Discussion

Our empirical analysis with simulated data suggested that Quor results are in line with the most frequently used techniques and might be a better option in some cases, for example when the data distribution is suspected to be non-Gaussian. The order of the groups that Quor provides as output might not be always available from hypothesis testing approaches, which is an extra benefit.

We also compared the methods in their ability to create an importance ranking of the variables. Using multiple real benchmark -omics data sets, we empirically verified that ranks obtained by Quor have a better association with univariate group prediction accuracy than those obtained with Student's *t* test, Mann–Whitney–Wilcoxon test, the modified *t* test from `samr`, and the Mood's median test. Using a copy number data set from DLBCL patients, we studied the use of Quor with quantiles other than the median, and empirically showed that opportune choices of quantiles based on domain knowledge and according to the particular analysis can produce more meaningful results. Quor obtained similar or better ranking than other approaches, including the Fisher exact test, which may be considered the most appropriate method for this type of data.

5. Conclusions

Quor represents a novel technique for ranking variables according to their relevance to predict a given outcome. It may perform better than state-of-the-art approaches depending on the characteristics of the problem in hand. This is evaluated in multiple scenarios with synthetic and real data.

As future work, we will further study the occurrence of numerical ties in the results of Quor. We consider it important to have a sound procedure to resolve them in order to deal with discrete ordinal data. We are also pursuing new applications of the method in biological data sets [25], in particular to the analysis of data with multiple groups and situations where some particular order of the groups is of great interest and importance.

Acknowledgments: This work was partially supported by the Brazilian agency CNPq grants 304025/2013-5 and 308776/2014-3. The authors thank Enrico Colossimo, who saw an initial analysis with the proposed method and encouraged us to write this paper. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions: All authors have participated in the design of the approach. Carlos A. de B. Pereira and Adriano Polpo have developed the statistical motivation and analysis. Cassio P. de Campos and Paola M. V. Rancoita have implemented and optimized the methods. All authors have participated in the analysis of results and in the writting of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Supporting Information

Software Package

The project named Quor is available at CRAN *http://cran.r-project.org/web/packages/Quor/*. It requires the programming languages `R` and `C++` and is distributed under the license GPL version 3.

Data Availability

The data sets supporting the results of this article are available (some by request, others available directly through the links) in the following repositories:

1. Kent Ridge Bio-medical Dataset
   http://datam.i2r.a-star.edu.sg/datasets/krbd/BreastCancer/BreastCancer.zip
   (files *breastCancer-train.arff* and *breastCancer-test.arff* combined),

   http://datam.i2r.a-star.edu.sg/datasets/krbd/NervousSystem/NervousSystem.zip
   (file *centralNervousSystem-outcome.arff*),

   http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Ontario.zip
   (file *lungcancer-ontario.arff*),

   http://datam.i2r.a-star.edu.sg/datasets/krbd/ProstateCancer/ProstateCancer.zip
   (file *prostate_outcome.arff*),

   http://datam.i2r.a-star.edu.sg/datasets/krbd/OvarianCancer/Ovarian-PBSII-061902.zip
   (file *ovarian_61902.arff*),
2. BioInformatics Group Seville
   http://eps.upo.es/bigs/dataSet/
   (files *leukemia_train_38x7129.arff* and *leukemia_test_34x7129.arff* combined),

   http://eps.upo.es/bigs/dataSet/
   (file *colon.arff*),
3. the Stanley Medical Research Institute genomics database
   https://www.stanleygenomics.org/ (full schizophrenia data set, upon request),
4. the Progenetix database
   (*http://www.progenetix.org*, array series ID: GSE11318).

References

1. Nguyen, D.V.; Rocke, D.M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **2002**, *18*, 39–50.
2. Raftery, A.E. Bayesian model selection in social research. *Sociol. Methodol.* **1995**, *25*, 111–164.
3. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
4. Mann, H.B.; Whitney, D.R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60.
5. Fay, M.P.; Proschan, M.A. Wilcoxon-Mann-Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat. Surv.* **2010**, *4*, 1–39.
6. Hayter, A.J. Simultaneous Confidence Intervals for Several Quantiles of an Unknown Distribution. *Am. Stat.* **2014**, *68*, 56–62.
7. Basu, D. On ancillary statistics, pivotal quantities and confidence statements. In *Topics in Applied Statistics*; Concordia University Publication: Montreal, QC, Canada, 1981; pp. 1–29.
8. Kiefer, J. Conditional confidence statements and confidence estimators. *J. Am. Stat. Assoc.* **1977**, *72*, 789–808.
9. Zellner, A.; Keuzenkamp, H.; McAleer, M. (Eds.) *Simplicity, Inference and Modeling: Keeping it Sophisticatedly Simple*; Cambridge University Press: Cambridge, UK, 2004.

10. DeGroot, M. *Probability and Statistics*, 2nd ed.; Addison-Wesley: New York, NY, USA, 1975.
11. Tusher, V.G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 5116–5121.
12. Wilcox, R.R.; Erceg-Hurn, D.M.; Clark, F.; Carlson, M. Comparing two independent groups via the lower and upper quantiles. *J. Stat. Comput. Simul.* **2014**, *84*, 1543–1551.
13. Wasserstein, R.L.; Lazar, N.A. The ASA's statement on *p*-values: Context, process, and purpose. *Am. Stat.* **2016**, *70*, 129–133.
14. Van't Veer, L.J.; Dai, H.; van de Vijver, M.J.; He, Y.D.; Hart, A.A.; Mao, M.; Peterse, H.L.; van der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415*, 530–536.
15. Pomeroy, S.L.; Tamayo, P.; Gaasenbeek, M.; Sturla, L.M.; Angelo, M.; McLaughlin, M.E.; Kim, J.Y.H.; Goumnerova, L.C.; Black, P.M.; Lau, C.; et al. Prediction of Central Nervous System Embryonal Tumour Outcome based on Gene Expression. *Nature* **2002**, *415*, 436–442.
16. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750.
17. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **1999**, *286*, 531–537.
18. Wigle, D.A.; Jurisica, I.; Radulovich, N.; Pintilie, M.; Rossant, J.; Liu, N.; Lu, C.; Woodgett, J.; Seiden, I.; Johnston, M.; et al. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.* **2002**, *62*, 3005–3008.
19. Petricoin, E.F., III; Ardekani, A.M.; Hitt, B.A.; Levine, P.J.; Fusaro, V.A.; Steinberg, S.M.; Mills, G.B.; Simone, C.; Fishman, D.A.; Kohn, E.C.; et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **2002**, *359*, 572–577.
20. Singh, D.; Febbo, P.G.; Ross, K.; Jackson, D.G.; Manola, J.; Ladd, C.; Tamayo, P.; Renshaw, A.A.; D'Amico, A.V.; Richie, J.P.; et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **2002**, *1*, 203–209.
21. Higgs, B.; Elashoff, M.; Richman, S.; Barci, B. An online database for brain disease research. *BMC Genom.* **2006**, *7*, 70.
22. Lenz, G.; Wright, G.W.; Emre, N.C.T.; Kohlhammer, H.; Dave, S.S.; Davis, R.E.; Carty, S.; Lam, L.T.; Shaffer, A.L.; Xiao, W.; et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13520–13525.
23. Baudis, M.; Cleary, M.L. Progenetix.net: An online repository for molecular cytogenetic aberration data. *Bioinformatics* **2001**, *17*, 1228–1229.
24. Nogai, H.; Dörken, B.; Lenz, G. Pathogenesis of Non-Hodgkin's Lymphoma. *J. Clin. Oncol.* **2011**, *29*, 1803–1811.
25. Bastos, E.P.; Brentani, H.; Pereira, C.A.B.; Polpo, A.; Lima, L.; Puga, R.D.; Pasini, F.S.; Osorio, C.A.B.T.; Roela, R.A.; Achatz, M.I.; et al. A Set of miRNAs, Their Gene and Protein Targets and Stromal Genes Distinguish Early from Late Onset ER Positive Breast Cancer. *PLoS ONE* **2016**, *11*, e0154325.