Knowledge and Information Systems (2022) 64:1549–1570
https://doi.org/10.1007/s10115-022-01677-6

REGULAR PAPER

Are the statistical tests the best way to deal with the biomarker selection problem?

Ari Urkullu[1] · **Aritz Pérez**[2] · **Borja Calvo**[1]

Received: 18 June 2018 / Revised: 20 March 2022 / Accepted: 21 March 2022 /
Published online: 8 May 2022

Abstract
Statistical tests are a powerful set of tools when applied correctly, but unfortunately the extended misuse of them has caused great concern. Among many other applications, they are used in the detection of biomarkers so as to use the resulting *p*-values as a reference with which the candidate biomarkers are ranked. Although statistical tests can be used to rank, they have not been designed for that use. Moreover, there is no need to compute any *p*-value to build a ranking of candidate biomarkers. Those two facts raise the question of whether or not alternative methods which are not based on the computation of statistical tests that match or improve their performances can be proposed. In this paper, we propose two alternative methods to statistical tests. In addition, we propose an evaluation framework to assess both statistical tests and alternative methods in terms of both the performance and the reproducibility. The results indicate that there are alternative methods that can match or surpass methods based on statistical tests in terms of the reproducibility when processing real data, while maintaining a similar performance when dealing with synthetic data. The main conclusion is that there is room for the proposal of such alternative methods.

Keywords Biomarker selection · Statistical tests · Reproducibility · Differential methylation detection

1 Background

In the first quarter of the twentieth century, Fisher popularized the concepts of probability and statistics in the field of scientific research. Besides developing the basic aspects of experimental design, Fisher worked on the problem of hypothesis testing and introduced the well-known (but poorly understood) concept of *p*-value.

✉ Ari Urkullu
ari.urkullu@ehu.eus

[1] Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Paseo Manuel de Lardizabal, 1, 20018 Donostia, Gipuzkoa, Spain

[2] Department of Data Science, Basque Center for Applied Mathematics BCAM, Alameda Mazarredo, 14, 48009 Bilbao, Bizkaia, Spain

Fisher's approach to hypothesis testing was that of a scientist, where one hypothesis is put to the test and further research steps depend on the result of that test. In such a scenario, the *p*-value was conceived as an indicator of the degree of evidence against the null hypothesis under test according to the gathered data [11].

Neyman and Pearson had a different view of the problem, one that matched the work of engineers, rather than scientists. In their approach, tests are used repeatedly to test a certain hypothesis so that the production parameters are within certain limits [20, 21]

There is a fundamental difference between both approximations; while in the view of Neyman and Pearson, the tests are used to make decisions, Fisher's tests are used as a guide in the scientific search quest. More importantly, the repeated use of the test gives meaning to concepts such as type I and type II errors, which are hardly interpretable when a test is run just once.

Unfortunately, at some point both approaches got mixed, a mix that became the null hypothesis significance testing (NHST). Briefly, NHST is a vaguely defined procedure that borrows concepts from those two different and rather incompatible approaches [7, 12, 22, 23]. Moreover, these tests seem to have reached, in many fields, an indisputable position [7, 29], from which the test that produces a *p*-value below a certain threshold[1] seems to grant a quality certificate that makes the result and the conclusions fully trustworthy [2, 7, 13, 16, 22, 29].

In addition, the many different misuses and misconceptions regarding the statistical tests [2, 7, 12, 13, 22, 29] seem to be quite extended [7]. Among others, it is worth mentioning the incorrect interpretations of the *p*-value [12, 13], the ignorance of the assumptions made by the given statistical test [13] or even the "p-hacking" [2, 16, 22], which consists of the reanalysis of the data in different forms until it yields a desired result.

The conjunction composed of the blind faith in the *p*-values and the misconceptions of the statistical tests has been identified as one of the main contributors to what has been called the reproducibility crisis [2, 29]. There are so many published conclusions drawn based on *p*-values that are false that the credibility of the scientific method is at risk [4]. The problem is so serious that some journals of areas such as psychology have even banned the use of *p*-values [13, 26, 30] and, eventually, the American Statistical Association (ASA), in answer to the long claim of many statisticians, has issued a warning about their misuses [29]. Any crisis brings changes, and, as many authors claim[2] [29], it is time to rethink how scientific results are analyzed and interpreted.

In this work, we focus on a popular problem in bioinformatics, the biomarker discovery in high-throughput data, in which mainly the statistical tests are used [14]. In this problem, the classic approach starts with the measurement of tens or hundreds of thousands of values (e.g., gene expression or methylation levels) in a few (hundreds or thousands) individuals from at least two populations (e.g., cases and controls) [1, 15].

A very frequent way to analyze the resulting matrices is to use some statistical tests to assess the degree of evidence against the null hypotheses, which mainly consist of no differences between the populations under analysis [14]. So, given that thousands of tests are run, some mathematical transformations are applied to the resulting *p*-values to cope with the problem of multiple testing [14]. Finally, among the n candidate biomarkers analyzed, the ones for which the corrected *p*-value is below a given threshold, say the traditional gold standard 0.05, are declared to behave differently depending on the population of individuals

[1] The threshold 0.05 is generally taken as the gold standard.

[2] The statement made by the ASA has been published along with the comments of a group of experts to whom a copy of the statement was sent prior to publication.

[14]. The corrected p-values are then interpreted in many incorrect ways, such as (1 minus) the probability of the null hypothesis being false [7, 12, 13, 22, 29] or the probability of obtaining the same results in a replication of the experiment [7, 12].

Interestingly, regardless of all those interpretations [13], it is generally accepted that the candidate biomarkers that obtained corrected p-values below the given threshold, though promising, are still candidate biomarkers that have yet to be validated [19]. That is, further validations are the only way to give credibility to the results. Certainly, these high-throughput experiments should be viewed as screening studies where promising candidate biomarkers are identified, promising candidate biomarkers that need to be properly validated in later studies from a biological perspective.

Now the question is are we forced to use statistical tests for this kind of analysis? Furthermore, if all the candidate biomarkers identified as promising (and not just some random ones) have yet to be validated, should the number of promising candidate biomarkers be determined by a hard-to-interpret probability (the p-value) and an arbitrary threshold, or by the available resources?

If the workflow to follow includes the validation of every promising candidate biomarker in later studies from a biological perspective, the first step should be to rank the candidate biomarkers according to the expected chances of confirming the observed differences in future studies (i.e., the reproducibility of the results). Even if such chances may be glimpsed through the degrees of evidence against the null hypotheses, the p-values tell very little about the reproducibility of results [2, 8]. Anyway, this step of generating a ranking of candidate biomarkers is by no means a new idea [14], as many studies end up with, among other results, a ranking of candidate biomarkers according to the corrected p-values of the given applied statistical test.

Consequently, if the final goal of the analysis is just ranking the candidates and not selecting a certain subset, there is no need to compute any p-value (corrected or not), as directly using the statistic defined in the test yields exactly the same ranking as both the p-values and the corrected p-values.

Moreover, the use of statistical tests to rank candidate biomarkers implies a set of constraints. Namely, for a given statistical test, the distribution of its statistic under its null hypothesis needs to be known so as to be able to compute p-values, which narrows the amount of eligible statistics. However, since there is no obligation to use methods that satisfy such a constraint for the task of ranking candidate biomarkers, there is way more freedom to devise alternative methods. Such freedom provides flexibility for the development of methods that can be as specific and ad hoc, or conversely as general, as needed. Consequently, bearing in mind such freedom, the question is—can we use alternative simple and intuitive methods to get good rankings? Moreover, how can we assess the goodnesses of both statistical tests and alternative methods, taking into account both their performance and their reproducibility? Therefore, in this article we make two contributions, consisting of the proposal of two alternative methods and the proposal of an evaluation framework, to answer each of those two questions.

In this work, in order to compare different methods that deal with the problem of the discovery of biomarkers in high-throughput data, we propose a general workflow, which is composed of several steps of increasing complexity, for the evaluation of such methods. Within that workflow, a combination of the ability to build rankings with synthetic data and the degree of reproducibility in real data are gathered as an assessment of the given methods.

Later, to illustrate that it is possible to design simple and intuitive alternative methods, which are able to build rankings as good (or better) as those achieved with statistical tests, the specific problem of differential methylation biomarker detection is tackled. We have chosen

this problem of the differential methylation biomarker detection because nowadays it is a popular topic in biology [24]. Within such problem, we compare several methods based on statistical tests with two alternative methods we have designed.

2 Methods

We have divided this section into two parts, one dedicated to the description of the general workflow and the other focused on the exposition of a particularization of this workflow for the differential methylation biomarker detection problem. In the first part, a general strategy to evaluate both the performance (in synthetic data) and the reproducibility (in real data) of any given ranking-based candidate biomarker selection (RCBS) method is exposed, together with the description of the evaluation measures used to assess the goodnesses of the results of the different RCBS methods compared. The second part illustrates a particularization of the process with an example, the analysis of RCBS methods applied to the differential methylation biomarker detection problem. Besides, within the second part, four RCBS methods based on statistical tests are presented. In addition, two alternative RCBS methods that are not based on statistical tests are proposed and described. Finally, the different approaches are compared using the proposed workflow.

2.1 General workflow for the evaluation of RCBS methods

In the general workflow to evaluate RCBS methods, we differentiate between two parts, namely the performance evaluation with synthetic data (Sect. 2.1.1) and the reproducibility evaluation with real data (Sect. 2.1.2). The use of synthetic data enables the performance of the RCBS methods to be evaluated under controlled circumstances. However, when using real data, unfortunately we do not know which candidate biomarkers really have differences or not, namely, we do not know which candidate biomarkers are true biomarkers or not. Consequently, the estimation of the performance with real data can hardly be made. Instead, our purpose is to evaluate the reproducibility of the results using real data.

2.1.1 Performance evaluation of RCBS methods with synthetic data

The part of the general workflow dedicated to the estimation of the performance using synthetic data is subdivided into three different stages. These three stages have multiple properties in common:

- All the data are sampled from known probability distributions.
- There are two populations of individuals (e.g., cases and controls).
- From one population a sample of M individuals is drawn, while from the other population a sample of N individuals is drawn.
- There is a subset of candidate biomarkers that are sampled from the same probability distribution for the two samples of individuals (non-biomarkers), while the rest of candidate biomarkers are sampled from different probability distributions for the different samples of individuals (true biomarkers).
- The nature of the differences that the candidate biomarkers present between populations (if they present any differences) is known beforehand (e.g., differences in centrality or differences in spread).

In the **first stage**, the motivation is to make a raw and simplified approximation to the real problem at hand. So, the artificial data are generated sampling just two certain probability distributions. Therefore, the comparison between populations is **binary**, i.e., two types of candidate biomarkers are distinguished in terms of differences, those for which there exists a difference between populations (the individuals of each sample have been drawn from different distributions), the true biomarkers, and those for which there are no differences between populations (the individuals of both samples have been drawn from the same distribution), the non-biomarkers. Besides, the distributions are simple parametric models, such as normal or beta distributions.

Regarding the evaluation of the results of the **first stage**, it is convenient to recall that after drawing the samples and applying a given RCBS method, a ranking of the candidate biomarkers is obtained. In addition, it is known whether each of those ranked candidate biomarkers is a biomarker or not. As any reordering of the candidate biomarkers of the same type is irrelevant[3], we propose the use of the AUC (area under the ROC curve) to evaluate such ranking in terms of how likely a biomarker will be before a non-biomarker in the ranking [10]. For instance, in the case of a RCBS method that perfectly ranks both types, namely ranking all the true biomarkers before all the non-biomarkers, the AUC will be 1, while for a RCBS method that ranks them randomly, the AUC will be 0.5 on average.

In the **second stage**, the motivation is to introduce the magnitude of the difference in play. Consequently, the complexity is increased and passes from a binary behavior of the candidate biomarkers to a **multilevel** one, i.e., now they are capable of showing several sizes of differences between populations. This change implies that now there are more than just two types (non-biomarkers and true biomarkers) of candidate biomarkers in terms of differences. Intuitively, it can be considered that the bigger the differences between populations for a given candidate biomarker, the higher the degree of its relevance.

In the ranking yielded by the given RCBS method in this second stage, there are more than just two types of candidate biomarkers. Therefore, now the AUC cannot be straightforwardly used anymore for evaluation purposes. Consequently, we propose the use of Kendall's τ distance, as it is a natural extension of the AUC [17] that enables the evaluation of rankings with more than just two types of candidate biomarkers. For instance, in the case of a RCBS method that perfectly ranks all the types[4], the Kendall's τ distance will be 0, while for a RCBS method that ranks them randomly, the Kendall's τ distance will be 0.5 on average.

In the **third stage**, we aim to make the synthetic data more realistic. In particular, we give room to situations in which the populations are not "homogeneous", such as when the populations are composed of subsets of individuals or when outliers are present. So, the complexity is increased again by sampling the values from distributions that are "heterogeneous", namely, multimodal distributions.

The evaluation of the results of the **third stage** is also conducted using Kendall's τ distance.

2.1.2 Reproducibility evaluation of RCBS methods with real data

The part of the workflow dedicated to the estimation of the reproducibility using real data is composed of a single stage. In this **fourth stage**, we use the general workflow presented in a recently published paper [27]. Briefly, the idea presented in that paper is composed of

[3] The differences between the distributions of the populations they have been sampled from are the same.

[4] Namely, all the candidate biomarkers of the most relevant type are ranked first, all the candidate biomarkers of the remaining most relevant type are ranked right after, and so on until all the candidate biomarkers of the least relevant type are ranked last.

several steps, an idea in which, first, two sets of data are sampled from a given original real dataset. Secondly, a ranking-based feature selection method under analysis is applied to those two sets of data, thus producing two different rankings of the n features of the given problem under study. Thirdly, the consistency between the two rankings is quantified so as to assess the reproducibility. Such consistency is quantified through the use of the consistency index developed by Kuncheva [18]. Specifically, Kuncheva's consistency index is applied to each possible pair of equally i-sized sets of top-ranked candidate biomarkers from the two rankings, thus obtaining a vector of n consistency indexes, for $i \in 1, \ldots, n$. The sequence $1, \ldots, n$ (in the abscissa axis) and the computed vector of n consistency indexes (in the ordinate axis) compose the curve (referred to as the reproducibility curve) that is the outcome of the third step.

The three aforementioned steps are run several times, and the different assessments of the reproducibility obtained in different runs are used to compute an average. Namely, the different obtained reproducibility curves are used to compute an estimation of the expected reproducibility curve.

Finally, a theoretical model for the reproducibility in ranking-based feature selections is fitted to the estimation of the expected reproducibility curve so as to gain insights, through the parameters of the fitted model, about how many relevant features there are and how relevant they are. For more details, see the aforementioned paper [27]. Briefly, the main outcome of the model is a sequence of n Booleans, identifying for each position of the sequence whether it issues a true biomarker or a non-biomarker. In addition, it also issues a weight which assesses the tendency to issue first (in the first positions of the ranking) true biomarkers to the detriment of non-biomarkers.

Specifically, in the aforementioned paper, the described procedure which, departing from a given real dataset, obtains a fit of the model is carried out twice, applying each time a different partition strategy to the original dataset. One of the partition strategies consists of dividing the original dataset into two equally sized subsets. This partition strategy has the characteristic of deriving two disjoint sets of half the size of the original real dataset, more than likely leading to a pessimistic estimation of the reproducibility. The other partition strategy consists of using a bootstrap scheme to derive the two sets. This partition strategy has the characteristics of deriving two sets that have instances in common and that are of the same size as the original real dataset, more than likely leading to an optimistic estimation of the reproducibility. This provides us with bounds for the true reproducibility.

2.2 Case of use: methylation biomarker ranking

In this subsection, the aforementioned general workflow is particularized for the differential methylation biomarker detection problem. In this problem in particular, the differential spread of methylation level can be as relevant as the differential centrality of the methylation level [6]. Indeed, some statistical tests have been specifically designed to be used in such situations [6]. So, in this experimentation, taking into account that property of the differential methylation biomarker detection problem, we compare between them statistical tests and alternative methods designed to detect differences only in centrality, and differences in both centrality and spread.

In the differential methylation biomarker detection problem, a standard way to output the DNA methylation level among the platforms based on microarray technologies consists of outputting β-values [9]. These β-values measure the relative amount of methylated molecules, expressing it within the range [0, 1]. Another standard way to express the DNA

methylation level is through M-values [9], which can be computed from the β-values through the use of the logit function, it being thus defined within the range $(-\infty, \infty)$ [9]. For this reason, in the stages in which synthetic data are used, we have considered mainly two families of probability distributions, the Beta and the Normal, so as to synthetically generate the β-values and the M-values, respectively. Further details about the experimentation can be found in the supplementary material (Online resource 1).

In Sects. 2.2.1, 2.2.2 and 2.2.3, we particularize the first three stages dedicated to evaluate the performance of RCBS methods with synthetic data. In Sect. 2.2.4, we particularize the last stage dedicated to evaluate the reproducibility of RCBS methods with real data. Finally, in Sect. 2.2.5, we present the six RCBS methods that are compared during the experimentation and we specify how two alternative RCBS methods which are not based on statistical tests that we propose can be computed.

2.2.1 First stage

In this stage, we have just 2 distributions (referred to as 'reference' and 'alternative'), 425 candidate biomarkers are drawn from the reference distribution for both samples, 425 candidate biomarkers are drawn from the only one alternative distribution available for both samples, and, finally, 150 candidate biomarkers are drawn using each distribution for one of the samples. This imbalance between the amounts of candidate biomarkers regarding the differences between populations follows what is expected to happen in high-throughput data in general. Namely, the amount of candidate biomarkers that show differences is expected to be much lower than the amount of candidate biomarkers that do not show differences [3]. This procedure is repeated for all the combinations of:

- Type of density function (Beta or Normal).
- Size of the samples ($20-20$, $50-50$ or $100-100$).
- Type of differences between the two distributions (centrality, or both centrality and spread).

For the sake of clarity, the sampling procedure is illustrated in Fig. 1.
Within each combination, 100 runs are carried out.

2.2.2 Second stage

In the second stage, instead of using just two distributions to generate the samples, we use five distributions. The two extremes are the same as in the previous stage and the other three are equally spaced between them (in terms of the centrality, or both the centrality or the spread).

All the candidate biomarkers are sampled from the reference distribution for one of the samples. For the other sample, the candidate biomarkers are sampled from the reference distribution (850 candidate biomarkers) and the other distributions (80, 40, 20 and 10 candidate biomarkers from the distribution with the smallest differences to the one with the largest differences, respectively).

Again, this procedure is repeated for all the combinations of type of distribution, sample size and type of difference. The sampling procedure is illustrated in Fig. 2.

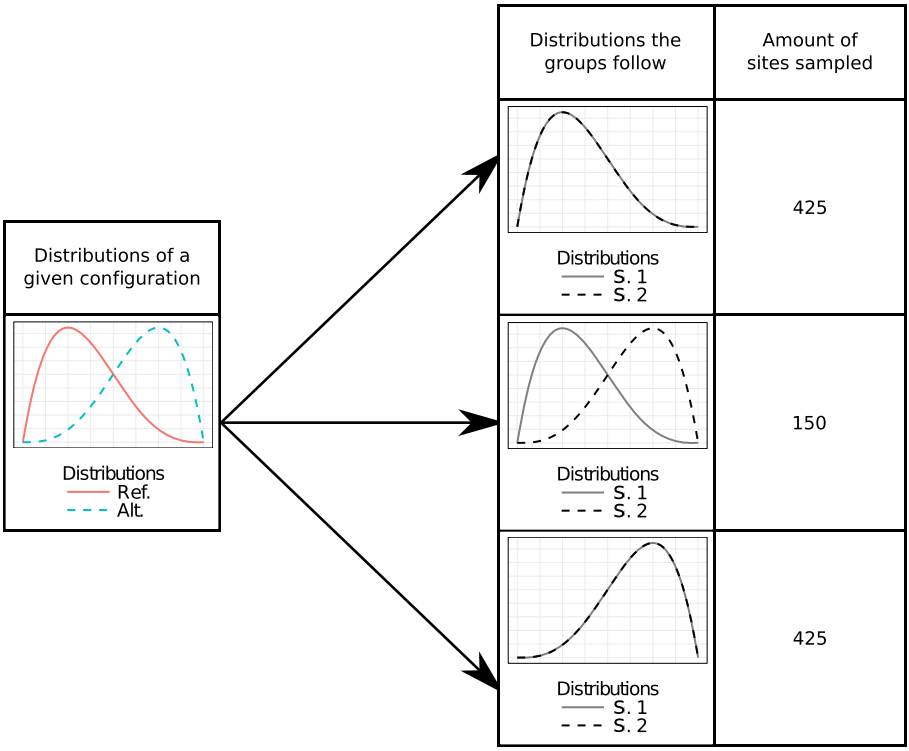


Fig. 1 Drawing procedure for the first stage for a given combination. Note that the distributions drawn in the figure are just used as an example and do not correspond to the actual distributions used in the experimentation

2.2.3 Third stage

The only change in this stage with respect to the previous one is that now the distributions are mixtures of two normal distributions. Three pairs of weights have been used for these mixtures, 0.5–0.5, 0.75–0.25 and 0.95–0.05. The latter case represents data with some degree of outliers.

For each pair of weights, the procedure followed is identical to that of the previous stage. Figure 2 also serves to illustrate the sampling procedure of this third stage, although that instead of using unimodal distributions, multimodal distributions are used.

2.2.4 Fourth stage

In this last fourth stage, we use real data to assess the reproducibility. Specifically, we have used two different real datasets (available at the GEO repository[5]). Each of them has two samples derived from two different populations (e.g., cases and controls) in which the methylation levels of 27578 CpG sites have been measured through the Illumina® Infinium® Human Methylation27 Beadchip technology.

[5] https://www.ncbi.nlm.nih.gov/geo/.

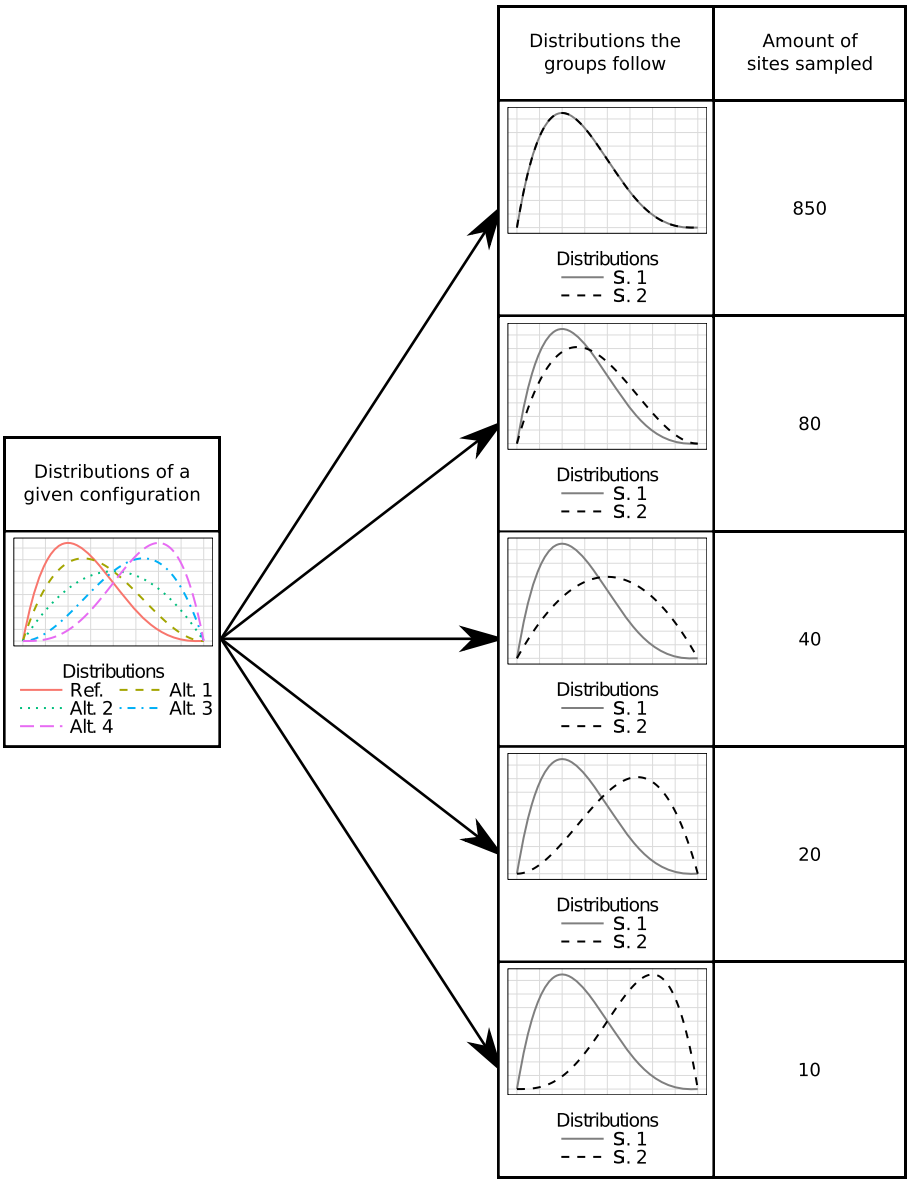


Fig. 2 Drawing procedure for the second and third stages for a given combination. Note that the distributions drawn in the figure are just used as an example and do not correspond to the actual distributions used in the experimentation

The first real dataset (accession number GSE19711[6]) [25] belongs to the United Kingdom Ovarian Cancer Population Study (UKOPS) and has been used in several papers [6, 25, 28].The dataset denotes through β-values the methylation level of the peripheral whole blood

[6] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19711.

of 540 postmenopausal women. Among them, there are 274 healthy controls and 266 ovarian cancer cases (131 pre-treatment cases and 135 post-treatment). The samples are age-matched, and the range of ages covered by the database is from 49 to 91 years.

The second dataset (accession number GSE20067[7]) [5, 25] belongs to a study of diabetic nephropathy in type 1 diabetes mellitus through an analysis of the DNA methylation [5, 25]. In this dataset, there are 195 individuals, 97 cases and 98 controls, including both men and women in an age range of 24 to 74 years old.

These datasets have been preprocessed according to the original papers [5, 25, 28]. The details of the preprocessings can be found in supplementary material (Online resource 1).

The reproducibility curves have been computed using two strategies to generate pairs of datasets (partition and bootstrap), and both are stratified in terms of size of the populations and general features (age, gender, etc.). Specifically, the whole strategy of sampling two datasets, processing them with the given RCBS method under analysis and computing the reproducibility curve, is run 10 times for each of the sampling procedures conceived (partition and bootstrapping). Finally, the 10 reproducibility curves are used to compute the estimation of the expected reproducibility curve to which the model is fitted right after.

2.2.5 Compared RCBS methods

One of the goals of this paper is showing that simple, more interpretable metrics can be used to rank biomarkers without the need for statistical tests. We will illustrate this in the context of methylation data.

A comparison will be conducted in two scenarios, one in which we are looking for differences in centrality and the other one in which we are looking for differences in both centrality and spread. For each case, we will compare three RCBS methods: a parametric (or semi-parametric) test, a nonparametric tests and an alternative simple metric. In particular:

- To identify differences in location only: T-test, as the parametric test; Wilcoxon test, as the nonparametric test; the Absolute Sum of the Differences (ASD) of the empirical cumulative distributions, as the alternative to the tests.
- To identify differences in location and spread: Tl test, a semiparametric test proposed by [6]; Kolmogorov–Smirnov as the nonparametric test; the Sum of the Absolute Differences (SAD) of the empirical cumulative distributions, as the alternative to the tests.

The two alternatives to the tests, ASD and SAD, share some calculations. For each site, the common part begins with the computation, for each sample, of the values of its empirical cumulative distribution ($\hat{F}_1$ and $\hat{F}_2$) at each of the values of **both** samples ($z_j \in G_1 \cup G_2$, $G_1 = \{x_1, \ldots, x_M\}$ and $G_2 = \{y_1, \ldots, y_N\}$). Next, the vector of one of the samples is subtracted from the other, thus obtaining the vector of signed differences:

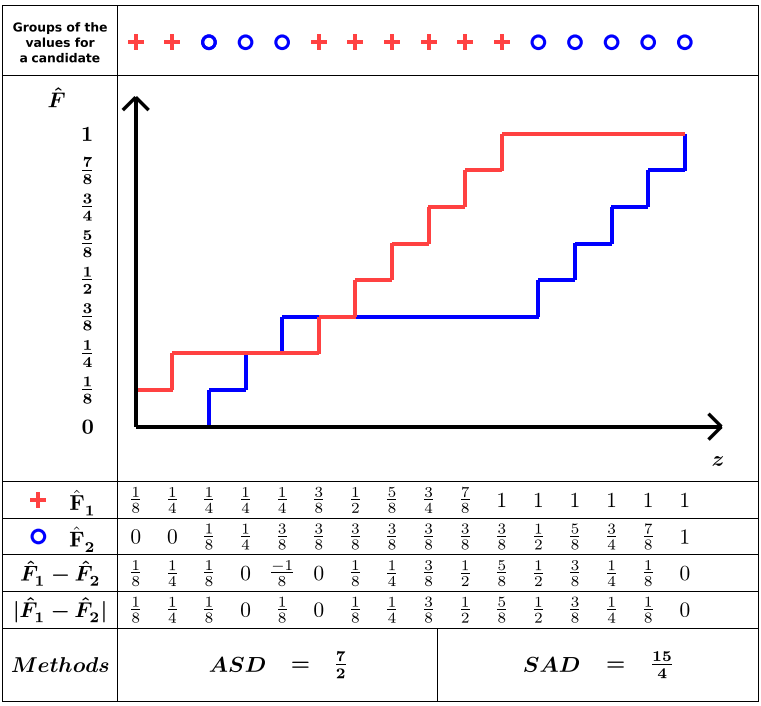
$$\hat{F}_1(z_j) - \hat{F}_2(z_j)$$

for $j \in 1, \ldots, M + N$. For each site:

- ASD: Sum all the differences and then get the absolute value:

$$\text{ASD} = \left| \sum_{j=1}^{M+N} \hat{F}_1(z_j) - \hat{F}_2(z_j) \right|$$

[7] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse20067



+ $\hat{F}_1$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1	1	1	1	1	1
o $\hat{F}_2$	0	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1
$\hat{F}_1 - \hat{F}_2$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{-1}{8}$	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0
$\lvert\hat{F}_1 - \hat{F}_2\rvert$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{1}{8}$	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0
Methods	*ASD* $= \frac{7}{2}$										*SAD* $= \frac{15}{4}$					

Fig. 3 Example 1: Computation of the two alternative methods in an example that fits in a scenario of differences in location

The intuition behind this method is that, as the differences in centrality become greater, it is more likely that more and more differences $\hat{F}_1(z_j) - \hat{F}_2(z_j)$ for $j \in 1, \ldots, M+N$ tend to be further away from zero and in the same direction (positive or negative). If the differences are only in spread, then this statistic will be close to 0.

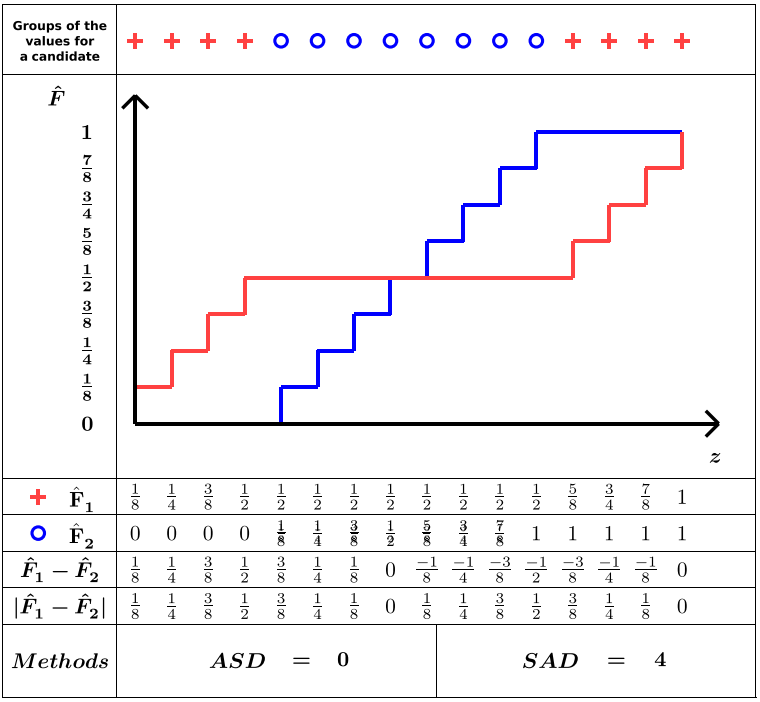
– SAD: Sum the absolutes of all the differences:

$$\text{SAD} = \sum_{j=1}^{M+N} \left|\hat{F}_1(z_j) - \hat{F}_2(z_j)\right|$$

The essence of this method lies in the fact that as the centralities and/or the spreads of the two distributions tend to differ, such differences tend to be portrayed in the terms $\left|\hat{F}_1(z_j) - \hat{F}_2(z_j)\right|$ for $j \in 1, \ldots, M+N$.

In order to clarify this explanation, the computation of the two alternative methods is shown for two different examples in Figs. 3 and 4.

It is worth mentioning that the two alternative methods designed and proposed, ASD and SAD, can be used in many different problems apart from the one in which they are used in this manuscript, the differential methylation biomarker detection problem. ASD and SAD do not make parametric assumptions and are not subject to the multiple conditions statistical tests have to meet. Therefore, it can be said that ASD and SAD are at least as flexible and as general as the statistical tests they are compared with.



+ $\hat{F}_1$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1
○ $\hat{F}_2$	0	0	0	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1	1	1	1	1	
$\hat{F}_1 - \hat{F}_2$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{-1}{8}$	$\frac{-1}{4}$	$\frac{-3}{8}$	$\frac{-1}{2}$	$\frac{-3}{8}$	$\frac{-1}{4}$	$\frac{-1}{8}$	0	
$\lvert\hat{F}_1 - \hat{F}_2\rvert$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0	
Methods	$ASD = 0$								$SAD = 4$								

Fig. 4 Example 2: Computation of the two alternative methods in an example that fits in a scenario of differences in spread

3 Results and discussion

In this section, the results of the different stages are presented and discussed.

Each configuration of the three synthetic stages has been run 100 times, and the performance measures are displayed in boxplots in Fig. 5 (differences in centrality) and Fig. 6 (differences in both centrality and spread). Specifically, the AUC is displayed for the first stage and $1 - \tau$ is displayed for the rest; consequently, in Figs. 5 and 6 the closer a boxplot is to 1, the better the performance.

Parametric methods are designed to deal with data sampled from a given type of statistical distribution. Consequently, it can be expected that the methods T-test and Tl test, which are parametric and semi-parametric, respectively, perform better when the synthetic data fit into their parametric or semi-parametric frames. That is, we can expect the T-test to achieve a better performance when dealing with data sampled from normal distributions, and the Tl test to achieve a better performance when dealing with data sampled from beta or normal distributions. Moreover, we can expect these methods to suffer the most with the data sampled from mixtures.

In the results of the first stage, it can be seen that, as expected, the parametric and semi-parametric tests behave better than the rest in the contexts they have been designed for. In addition, the T-test performs better than the Wilcoxon test and the ASD method even when the synthetic data are generated using beta distributions. When only differences in centrality are present, Wilcoxon and the ASD method behave very similarly. When differences both in centrality and spread are present, it appears that the SAD method achieves a better

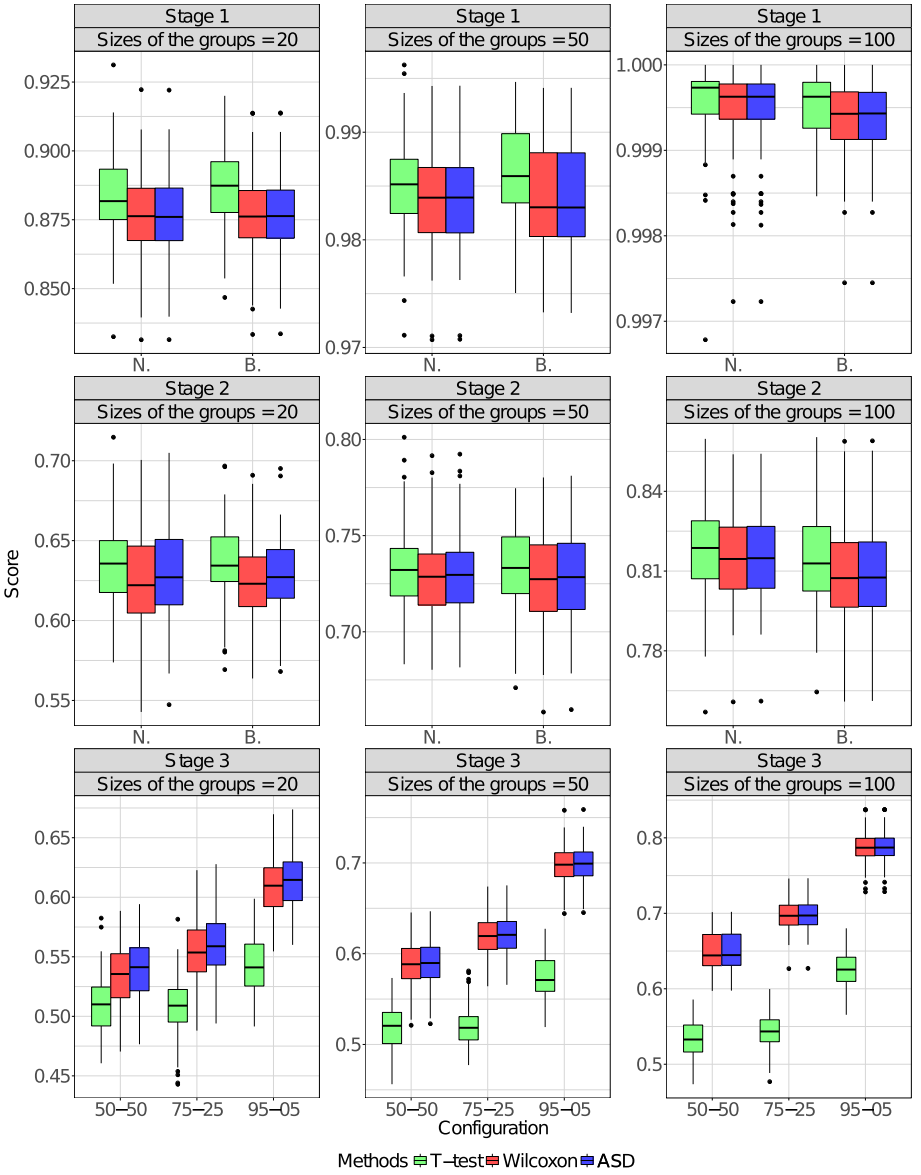


Fig. 5 Results in the synthetic stages of the methods for differences in centrality. The labels of the abscissa axes of the boxplots specify information about the distributions used: "N."—normal distributions, "B."—beta distributions, "50 − 50", "75 − 25" or "95 − 05"—mixtures of normal distributions in which the weights of the normal distributions are equal to the values specified by the corresponding label

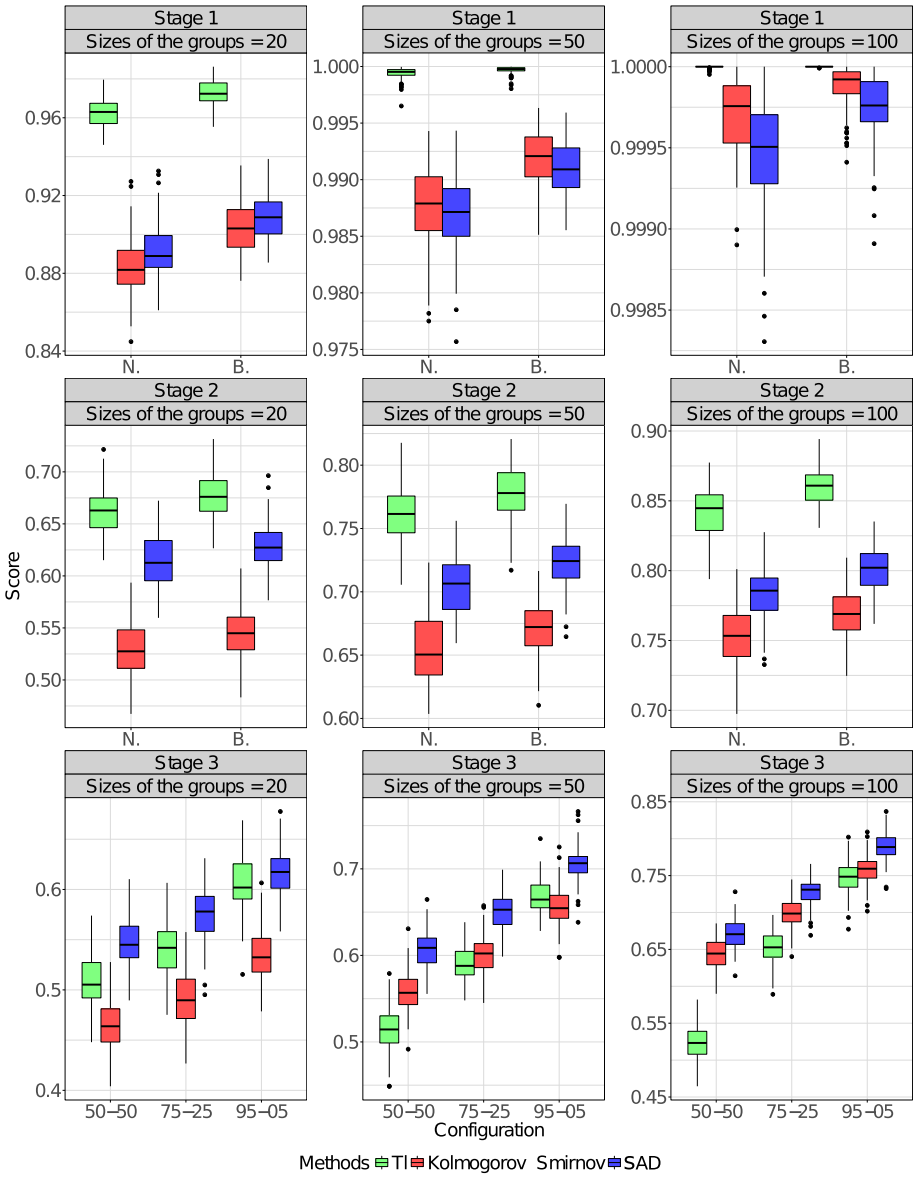


Fig. 6 Results in the synthetic stages of the methods for differences in centrality and spread. The labels of the abscissa axes of the boxplots specify information about the distributions used: "N."—normal distributions, "B."—beta distributions, "50 − 50", "75 − 25" or "95 − 05"—mixtures of normal distributions in which the weights of the normal distributions are equal to the values specified by the corresponding label

performance than the Kolmogorov–Smirnov test when the samples are small (20 individuals per sample) but worse when the samples are bigger (50 and 100 individuals per sample).

Regarding the results of the second stage, it can be seen that the performances of all the methods drop because the problem has become more difficult. Again, as expected, the parametric and semi-parametric tests behave better within their natural contexts. Also in this second stage, the T-test performs better than the Wilcoxon test and the ASD method when the synthetic data are generated using beta distributions. Besides, Wilcoxon and ASD keep obtaining similar performances. However, changes appear in the relative performance between the Kolmogorov–Smirnov test and the SAD method, this last one reaching better average results than the Kolmogorov–Smirnov test in every configuration of the second stage.

In the third stage, as expected, the performance of the (semi-)parametric methods drops in comparison with the other two. In the case of the T-test, it achieves worse performances than both the Wilcoxon test and the ASD method in every configuration of the third stage. However, the similarity in the performance between the Wilcoxon test and the ASD method still remains. In the scenario of differences in both centrality and spread, the SAD method achieves, in every configuration of the third stage, the best average performance, while the Tl test and the Kolmogorov–Smirnov test interchange positions between them depending on the specific configuration.

The results of experimentation with real data are shown in Fig. 7 and in Tables 1, 2, 3 and 4. This time, in Fig. 7, the x axis shows the logarithm of the sizes of the tops, while the y axis shows the estimations derived from the data of the n expected consistency indexes. Further details can be found in the aforementioned paper [27]. In each subfigure, for each one of the three methods compared, two different lines are shown. One corresponds to the estimation of the expected reproducibility when the original dataset given is partitioned into disjoint subsets (the pessimistic lower curve). The other one corresponds to the estimation of the expected reproducibility when the two new datasets are derived with a sampling through bootstrapping in the original dataset given (the optimistic higher curve). Namely, due to their nature, the two estimations have opposite biases, and thus, the true curve most likely lies somewhere between both. Finally, it is worth mentioning that the n expected consistency indexes of a method which assesses candidate biomarkers according to a random uniform distribution is 0.

The reproducibility displayed in the plots can be summarized in two values. Assuming the simplification of having only two types of candidate biomarkers (true biomarkers and non-biomarkers), we can estimate the number of candidate biomarkers with differences and their relative weight (the tendency to consistently appear in the first positions). These values are shown in Tables 1, 2, 3 and 4. Specifically, in each cell of the tables two values are displayed, the first one associated with the use of a stratified partition and the second one associated with the use of bootstrapping.

In the case of the ovarian cancer database, in general, the behaviors of the RCBS methods show, in Subfigures 7a and b, similar shapes that tend to follow a pattern. This pattern of the shapes first makes a hill-shaped curve moving away from the random behavior, through a steep slope until the top of the hill is reached. Then, its reproducibility decreases approaching the random behavior until finally the amount of candidate biomarkers in the top rankings is the whole set of candidate biomarkers. This general behavior matches a scenario in which the RCBS methods consistently assess a few candidate biomarkers as more relevant than the rest of the candidate biomarkers. Consequently, they tend to appear in the first positions of the rankings consistently, while the orders of the rest of candidate biomarkers are frequently interchanged by the RCBS methods. More specifically, the T-test shows curves of reproducibility that are in general lower than the reproducibility curves of the other two RCBS

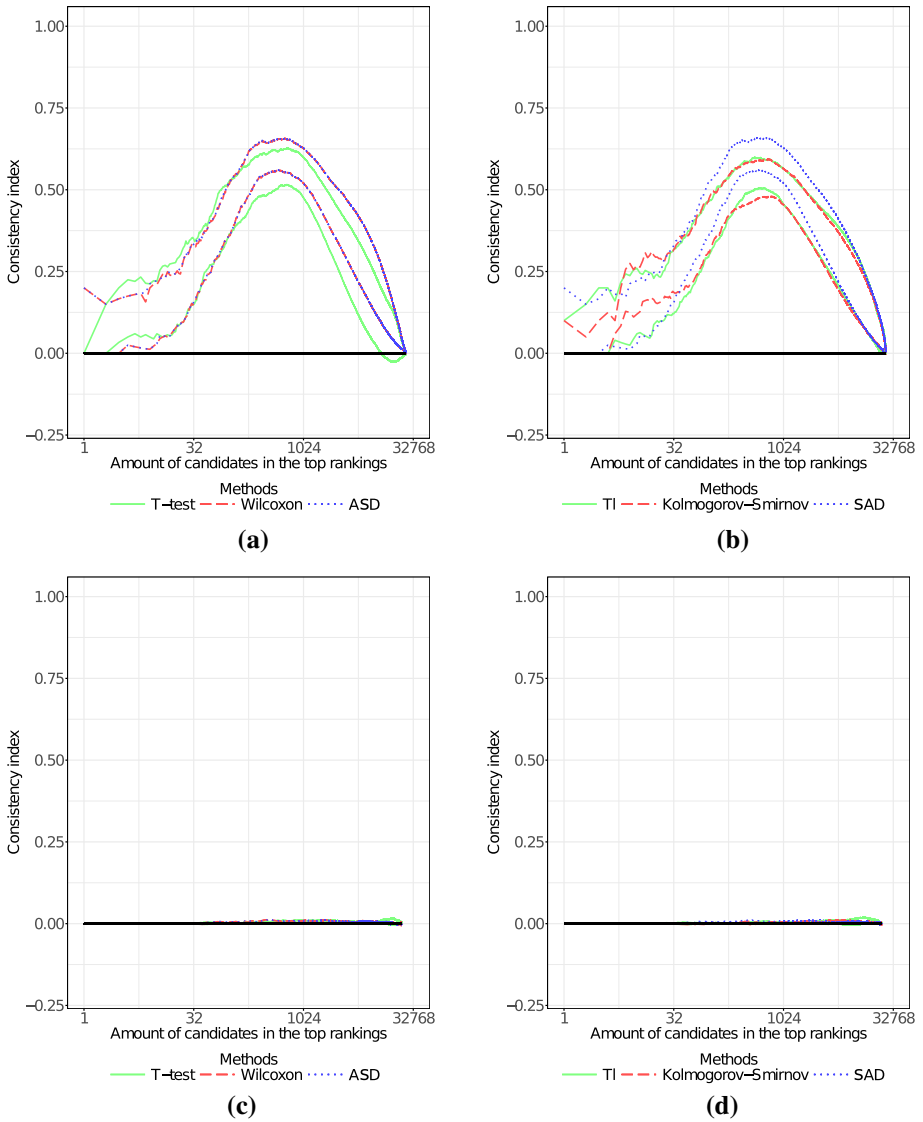


Fig. 7 Results in the ovarian cancer database, when 7**a** methods for differences in centrality are applied and 7**b** methods for differences in centrality and spread are applied; and results in the nephropathy database, when 7**c** methods for differences in centrality are applied and 7**d** methods for differences in centrality and spread are applied

methods. Besides, for the Wilcoxon test and the ASD method, the values that the fitted model issues offer a mixed outcome, issuing a lower amount of true biomarkers with greater weights for both of them than for the T-test. Regarding the comparison of the RCBS methods used to face the scenario with both differences in centrality and spread, in the ovarian cancer database the curves of SAD are in general higher than the curves of the other two RCBS methods. In terms of values issued by the fitted model, SAD issues weight values higher than the other two RCBS methods, but it issues an amount of true biomarkers comparable to the amount issued for the Kolmogorov–Smirnov test.

Table 1 Amounts and weights of candidate biomarkers with differences for the methods for differences in centrality when applied to the ovarian cancer database

Method	Amounts	Weights
T-test	(734, 2578)	(105.21, 26.34)
Wilcoxon	(1340, 3508)	(43.73, 24.02)
ASD	(1339, 3500)	(43.78, 24.09)

Table 2 Amounts and weights of candidate biomarkers with differences for the methods for differences in centrality and spread when applied to the ovarian cancer database

Method	Amounts	Weights
Tl	(1349, 3941)	(33.12, 17.15)
Kolmogorov-Smirnov	(1310, 3619)	(33.10, 17.55)
SAD	(1336, 3611)	(43.91, 21.90)

Table 3 Amounts and weights of candidate biomarkers with differences for the methods for differences in centrality when applied to the nephropathy database

Method	Amounts	Weights
T-test	(7459, 7780)	(0.78, 1.00)
Wilcoxon	(6956, 6809)	(1.04, 1.11)
ASD	(7661, 6596)	(1.04, 1.15)

Table 4 Amounts and weights of candidate biomarkers with differences for the methods for differences in centrality and spread when applied to the nephropathy database

Method	Amounts	Weights
Tl	(8680, 10288)	(0.87, 1.32)
Kolmogorov-Smirnov	(9127, 10075)	(1.03, 1.20)
SAD	(8065, 6395)	(1.15, 1.15)

In contrast, in the case of the nephropathy database, the behaviors of the RCBS methods show, in Subfigures 7c and d, a very different pattern. Specifically, the pattern consists of a totally flat curve, which matches a scenario in which all the RCBS methods fail to achieve a reproducibility any different from that of a random RCBS method. Namely, such behavior indicates that none of the RCBS methods can consistently assess any candidate biomarkers as any more relevant than any other. One possible explanation for this general behavior is that the differences between populations are so small that the RCBS methods are barely able to detect them. Subsequently, any change in the sample leads to changes in the rankings the RCBS methods produce. Another possibility is that there may be some problem with the data. Another possible explanation is that the RCBS methods do not show preferences towards any candidate biomarker, and therefore, they produce rankings at random. In addition, the values that the fitted model issues are clearly congruent, issuing for all the cases weights which are really close to one. Such an occurrence suggests that if there are any true biomarkers in the dataset, they do not tend to be drawn consistently before the non-biomarkers by the RCBS methods.

Finally, it is worthy mentioning that the results displayed in this paper are not the only ones generated by our research. In the supplementary material (Online resource 1), results corresponding to some other alternative methods that we studied are gathered.

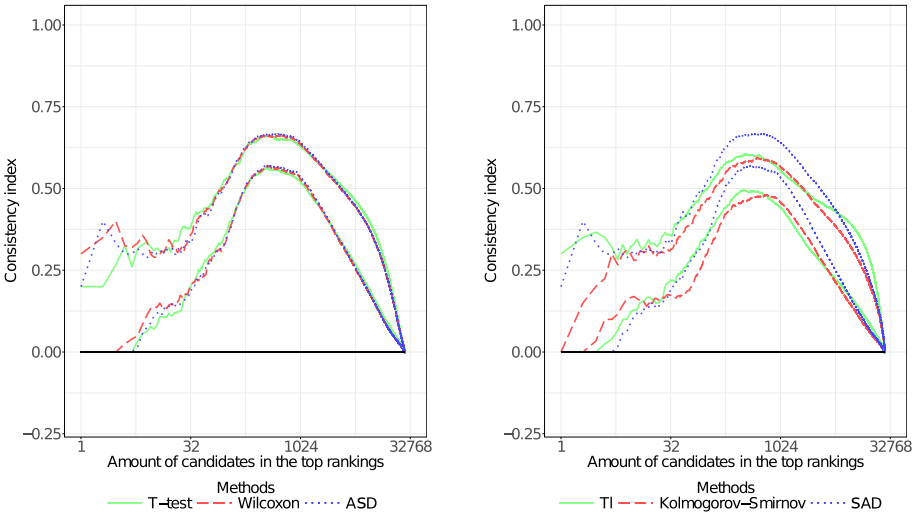


Fig. 8 Results in the ovarian cancer database when outliers are removed, when 8**a** methods for differences in centrality are applied and 8**b** methods for differences in centrality and spread are applied

Table 5 Amounts and weights of candidates with differences for the methods for differences in centrality when applied to the ovarian cancer database in which outliers have been removed

Method	Amounts	Weights
T-test	(1851, 4644)	(30.44, 21.06)
Wilcoxon	(1682, 4161)	(34.79, 22.33)
ASD	(1690, 4147)	(35.53, 22.63)

Table 6 Amounts and weights of candidates with differences for the methods for differences in centrality and spread when applied to the ovarian cancer database in which outliers have been removed

Method	Amounts	Weights
Tl	(1914, 6343)	(18.97, 13.40)
Kolmogorov-Smirnov	(1555, 4281)	(27.61, 16.01)
SAD	(1635, 4147)	(36.26, 22.63)

3.1 Analysis of the impact of the outliers

In addition to all the experimentation previously exposed, we decided to run an additional experiment so as to assess how much the outliers present in real databases impact the compared RCBS methods in terms of reproducibility. Specifically, we decided to rerun the experimentation carried out with the ovarian cancer database, modifying the preprocessing done. We added to it steps so as to systematically remove every single value from the data labeled as an outlier. Further details can be found in supplementary material (Online resource 1). This decision was based on two motivations: One is that we wanted to see how much the withdrawal of outliers in real data helps to increase the reproducibility of the RCBS methods that showed a worse performance when dealing with synthetic data in which outliers were present. The other motivation that encouraged us to rerun only the experimentation performed on the ovarian cancer database is that it is the experimentation with real data in which the different RCBS methods show different behaviors. The results corresponding to this additional experimentation can be seen in Fig. 8 and in Tables 5 and 6.

Analyzing the new results and taking into account the previous ones, in comparison with the RCBS methods for differences in centrality it can be seen that, as expected, the T-test method is that which benefits the most. In the new results, it achieves a reproducibility curve comparable to the ones of the other two RCBS methods for differences in centrality. Besides, some improvement can be seen in the curves of the three RCBS methods in the leftmost part (first positions) of them. In terms of the values issued by the fitted model, now the previous mixed outcome reverts, this time the T-test being the one with a greater amount of true biomarkers with lesser weight values.

Comparing the RCBS methods for differences in centrality and spread in terms of reproducibility curves, although it is not as clear as in the previous comparison, it seems that the Tl method benefits more than the other two. Again, all the methods improve their curves in the leftmost part. Additionally, in the case of the Tl method, it seems that its curve of reproducibility when a stratified partition is performed improves mainly by shifting leftwards. In addition, in the case of the reproducibility curve of the Tl method when bootstrap is done, it can be seen that it also improves in the most right-sided part. Besides, in terms of values issued by the fitted model, the amounts of true biomarkers issued for Tl increases proportionally more than it increases for the other two methods. In contrast, the weights of the Tl are the ones that drop more. In general, all the methods increase their amounts of true biomarkers while dropping their weights, except for the SAD method when bootstrapping is applied, which increases in both.

In global terms, considering the results of the analysis of the impact of the outliers and taking into account the previous results, it seems that the alternative RCBS methods show at least a better robustness to outliers in comparison with the RCBS methods based on parametric and semi-parametric statistical tests.

4 Conclusions

Since the computation of p-values is not necessary to rank candidate biomarkers, in this work we have explored mainly one question: Is it possible to design simple, alternative metrics to get better rankings? In order to answer this question, we have developed two alternative metrics and we have developed an evaluation workflow where both the performance and the reproducibility of any method can be assessed.

As an example, we have compared very simple RCBS methods with classical and state-of-the-art tests in the context of methylation data. From this comparison, we can draw some conclusions, and, more important, an answer to the main question. The results of the comparison show that, as expected, parametric models work great when their assumptions are true but, in more complex situations, such as multimodality or the presence of outliers, nonparametric methods work better. As for the proposed alternative RCBS methods not based on statistical tests, their behavior is similar to the RCBS methods based on nonparametric statistical tests in the case of differences in centrality, but shows a better compromise between different scenarios in the case of differences in both centrality and spread.

The plots used in the last stage of our workflow show in a simple, visual way the stability of the methods in a particular dataset. As such, they can be a good tool to analyze the RCBS methods or the effect of the preprocessing, as shown in the differences between the curves of the RCBS method based on the T-test when a more aggressive removal of outliers is conducted. Moreover, they can also be a great tool to analyze the data itself (in the case of nephropathy data, none of the methods is able to obtain consistent results, suggesting that

there may be a problem with the data itself). Regarding the reproducibility, we can see that the proposed alternative RCBS methods not based on statistical test are, in general, more stable than the other RCBS methods.

To sum up, going back to the main question, the results show that, indeed, we can design good, competitive methods to rank biomarkers without the restrictions imposed by the statistical tests. This freedom enables both the proposal of methods as specific as needed so as to cope with the particularities of any given problem at hand and the proposal of methods as general as desired for their application to a wide range of problems. For that task (proposing new metrics), evaluating the alternatives is absolutely mandatory. In this paper, we have shown how this comparison can be made, not only in the results of performance but also from the reproducibility point of view. Moreover, this analysis workflow can be used to evaluate other procedures (such as preprocessing steps or strategies to increase the robustness of the analysis) in terms of reproducibility.

Finally, we would like to outline some possible future work research lines. One possible work line could be the development and testing of strategies to increase the robustness of the methods used to rank biomarkers. Another possibility consists of the extension of the workflow through the introduction of a new stage in which the RCBS methods would have to deal with data generated through simulation. Briefly, in contrast to the synthetic stages in which too simplistic distributions are used, in this new stage the biology underlying the given specific problem at hand would be simulated. This way, the performance of the RCBS methods could be assessed in scenarios far more similar to the real problem under study than the scenarios in which synthetic data have been used. Lastly, the design of new specific RCBS methods for particular problems is another research line that should be taken into account.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10115-022-01677-6.

Acknowledgements This work is partially supported by the Basque Government (IT1244-19, Elkartek BID3A and Elkartek project 3KIA, KK2020/00049) and the Spanish Ministry of Economy and Competitiveness MINECO (PID2019-104966GB-I00) and a University-Society Project 15/19 (Basque Government and University of the Basque Country UPV/EHU). Ari Urkullu has been supported by the Basque Government through a predoctoral grant (PRE_2013_1_1313, PRE_2014_2_87, PRE_2015_2_0280 and PRE_2016_2_0314). Aritz Pérez has been supported by the Basque Government through the BERC 2022-2025 and Elkartek programs and by the Ministry of Science, Innovation and Universities: BCAM Severo Ochoa accreditation SEV-2017-0718. Borja Calvo has been partially supported by the IT1244-19 project and the ELKARTEK program from Basque Government, and the project PID2019-104966GB-I00 from the Spanish Ministry of Economy and Competitiveness.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alzubaidi AHA (2019) Evolutionary and deep mining models for effective biomarker discovery. Nottingham Trent University (United Kingdom)
2. Amrhein V, Korner-Nievergelt F, Roth T (2017) The earth is flat (p 0.05): significance thresholds and the crisis of unreplicable research. PeerJ 5:e3544
3. Baek S, Tsai CA, Chen JJ (2009) Development of biomarker classifiers from high-dimensional data. Brief Bioinform 10(5):537–546
4. Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature News 533(7604):452
5. Bell CG, Teschendorff AE, Rakyan VK, Maxwell AP, Beck S, Savage DA (2010) Genome-wide dna methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. BMC Med Genomics 3(1):1–11
6. Chen Y, Ning Y, Hong C, Wang S (2014) Semiparametric tests for identifying differentially methylated loci with case-control designs using illumina arrays. Genet Epidemiol 38(1):42–50
7. Cohen J (1995) The earth is round (p 05): Rejoinder
8. Colquhoun D (2017) The reproducibility of research and the misinterpretation of p-values. R Soc Open Sci 4(12):171085
9. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM (2010) Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. BMC Bioinform 11(1):587
10. Fawcett T (2006) An introduction to roc analysis. Pattern Recognit Lett 27(8):861–874
11. Fisher RA (1925) Statistical methods for research workers. Genesis Publishing Pvt Ltd, Edinburgh, London
12. Goodman S (2008) A dirty dozen: twelve p-value misconceptions. Semin Hematol 45(3):135–140
13. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31(4):337–350
14. Hahne F, Huber W, Gentleman R, Falcon S (2010) Bioconductor Case Studies. Springer Science & Business Media, New York city, New York
15. He Z, Yu W (2010) Stable feature selection for biomarker discovery. Comput Biol Chem 34(4):215–225
16. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. PLoS Biol 13(3):e1002106
17. Hernández-Orallo J, Flach P, Ferri C (2013) Roc curves in cost space. Mach Learn 93(1):71–91
18. Kuncheva LI (2007) A stability index for feature selection. In: Artificial intelligence and applications, pp 421–427
19. Lay JO Jr, Liyanage R, Borgmann S, Wilkins CL (2006) Problems with the "omics". TrAC Trends Anal Chem 25(11):1046–1056
20. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: part I. Biometrika 20A(1–2):175–240
21. Neyman J, Pearson ES (1933) The testing of statistical hypotheses in relation to probabilities a priori. In: Mathematical Proceedings of the Cambridge Philosophical Society, 29, pp 492–510. Cambridge University Press
22. Nuzzo R (2014) Statistical errors. Nature 506(7487):150–152
23. Perezgonzalez JD (2015) Fisher, neyman-pearson or nhst? a tutorial for teaching data testing. Front Psychol 6:223
24. Schübeler D (2015) Function and information content of dna methylation. Nature 517(7534):321–326
25. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP et al (2010) Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 20(4):440–446
26. Trafimow D, Marks M (2015) Editorial. Basic Appl Soc Psychol 37(1):1–2
27. Urkullu A, Pérez A, Calvo B (2021) Statistical model for reproducibility in ranking-based feature selection. Knowl Inf Syst 63(2):379–410
28. Wang S (2011) Method to detect differentially methylated loci with case-control designs using illumina arrays. Genet Epidemiol 35(7):686–694
29. Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. Am Stat 70(2):129–133
30. Woolston C (2015) Psychology journal bans P values. Nature 519(7541):9

Ari Urkullu received the MSc degree in computational engineering and intelligent systems from the University of the Basque Country in 2011. From 2011 to 2013, he was a software developer at Fullstep Networks. In 2014, he joined Intelligent Systems Group. From 2014 to 2018, he worked as a predoctoral researcher at the University of the Basque Country. In 2018, he was a temporary lecturer at the Department of Languages and Information Systems at the University of the Basque Country. From 2019 to the present, he has been working at Gestamp, where he currently works as a senior data scientist of the advanced analytics team. Besides, he works to finish his Ph.D. in informatics engineering under the supervision of Borja Calvo and Aritz Pérez. His research interests include bioinformatics, supervised and unsupervised classification, feature selection, model selection and evaluation, and both classification and forecasting of multivariate time series.

Aritz Pérez received his PhD degree in 2010 from the University of Basque Country, Department of Computer Science and Artificial Intelligence. Currently, he is a postdoctoral researcher at the Basque Center for Applied Mathematics. His current scientific interests include supervised, unsupervised and weak classification, probabilistic graphical models, model selection and evaluation, time series, and crowd learning.

Borja Calvo received his master's degree in Biochemistry from the University of the Basque Country in 1999, and, after two years working, he took a bachelor's degree in Computer Science at the same university. In 2004 he earned the bachelor's degree and in 2008 the PhD in computer science. After two years as a postdoc researcher at the Intelligent Systems Group, in 2011 he won his current lecturer position at the Department of Computer Science and Artificial Intelligence of the University of the Basque Country. Currently, he is leading a research project funded by the DGT (Spanish traffic agency) aimed at the prediction of car accidents in the Basque road network. He is also supervising three PhD students and several master thesis.