RESEARCH ARTICLE

Analyzing biomarker discovery: Estimating the reproducibility of biomarker sets

Amir Forouzandeh[1]*, Alex Rutar[2], Sunil V. Kalmady[1,3], Russell Greiner[1,4]

1 Department of Computing Science, University of Alberta, Edmonton, Canada, **2** Department of Pure Math, University of Waterloo, Waterloo, ON, Canada, **3** Canadian VIGOUR Centre, University of Alberta, Edmonton, Canada, **4** Alberta Machine Intelligence Institute, Edmonton, Canada

* forouzan@ualberta.ca

Abstract

Many researchers try to understand a biological condition by identifying *biomarkers*. This is typically done using univariate hypothesis testing over a labeled dataset, declaring a feature to be a biomarker if there is a significant statistical difference between its values for the subjects with different outcomes. However, such sets of proposed biomarkers are often not reproducible – subsequent studies often fail to identify the same sets. Indeed, there is often only a very small overlap between the biomarkers proposed in pairs of related studies that explore the same phenotypes over the same distribution of subjects. This paper first defines the *Reproducibility Score* for a labeled dataset as a measure (taking values between 0 and 1) of the reproducibility of the results produced by a specified fixed biomarker discovery process for a given distribution of subjects. We then provide ways to reliably estimate this score by defining algorithms that produce an over-bound and an under-bound for this score for a given dataset and biomarker discovery process, for the case of univariate hypothesis testing on dichotomous groups. We confirm that these approximations are meaningful by providing empirical results on a large number of datasets and show that these predictions match known reproducibility results. To encourage others to apply this technique to analyze their biomarker sets, we have also created a publicly available website, https://biomarker.shinyapps.io/BiomarkerReprod/, that produces these *Reproducibility Score* approximations for any given dataset (with continuous or discrete features and binary class labels).

Citation: Forouzandeh A, Rutar A, Kalmady SV, Greiner R (2022) Analyzing biomarker discovery: Estimating the reproducibility of biomarker sets. PLoS ONE 17(7): e0252697. https://doi.org/10.1371/journal.pone.0252697

Editor: Nguyen Quoc Khanh Le, Taipei Medical University, TAIWAN

Published: July 28, 2022

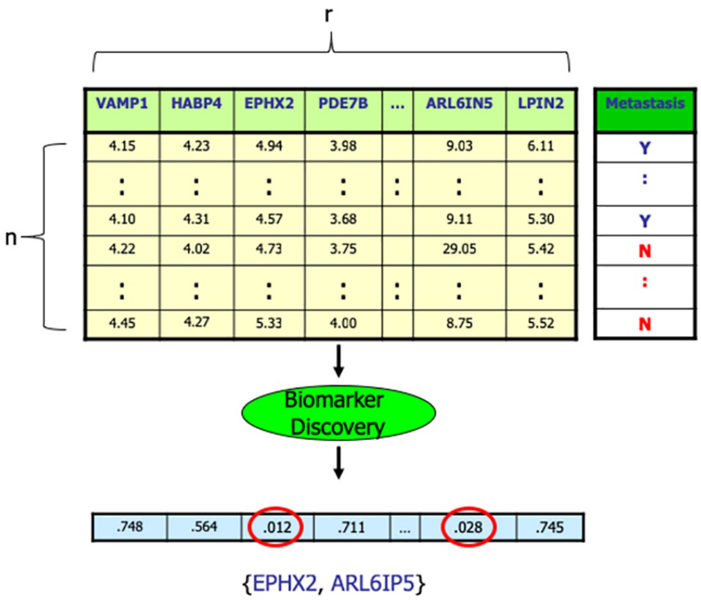
Data Availability Statement: All relevant data are within the manuscript and its Supporting information section and the URLs for accessing the data (all are public datasets) are as follows: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser https://ega-archive.org/datasets/EGAD00010000268 https://portal.gdc.cancer.gov/projects/TCGA-BRCA http://linkedomics.org/data_download/TCGA-KIPAN/.

Funding: The authors all gratefully acknowledge the funding support from Canada's NSERC (Natural Sciences and Engineer ResearchCouncil) and the Alberta Machine Intelligence Institute (Amii).

Competing interests: The authors declare no competing interests.

1 Introduction

Improved understanding of a disease can lead to better diagnosis and treatment. This often begins by finding "biomarkers", which is a generic term referring to "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"[1]. Typically, these are individual features (*e.g.*, expression values of specific genes [2, 3]) that follow different distributions (*e.g.*, have different mean values) in diseased patients versus healthy controls.



VAMP1	HABP4	EPHX2	PDE7B	...	ARL6IN5	LPIN2	Metastasis
4.15	4.23	4.94	3.98		9.03	6.11	Y
:	:	:	:	:	:	:	:
4.10	4.31	4.57	3.68		9.11	5.30	Y
4.22	4.02	4.73	3.75		29.05	5.42	N
:	:	:	:	:	:	:	:
4.45	4.27	5.33	4.00		8.75	5.52	N

.748	.564	.012	.711	...	.028	.745

Fig 1. Data matrix, showing t-test p-values for each (shown) feature for the GSE 7390 dataset [14], with respect to the group outcome (here "Metastasis" for breast cancer). The circled features, with $p<0.05$, are (purported) biomarkers. For notation: We will refer to each of the first *r* columns of the matrix as a "feature"; these are often called "(independent) variables". We refer to the final column as a "outcome"–*e.g.*, case versus control (shown here as Y versus N)–these are often called "labels", "dependent variables", "groups", "phenotypes" or "classes". Finally, we will use "subject" to refer to each row of that matrix; these are sometimes called "instances" or "samples".

https://doi.org/10.1371/journal.pone.0252697.g001

Sometimes, biomedical researchers can identify candidates for biomarkers based on their existing knowledge of the disease etiology and/or cellular pathways. This is done by seeking features that are causally related to the disease (*e.g.*, phenylketonuria is caused by mutations in the single gene PAH [4]) or a symptom of it (*e.g.*, Hemoglobin A1C for monitoring the degree of glucose metabolism in diabetes [5]). This paper, however, focuses on the use of statistical tools to discover and evaluate biomarkers, which is typically based on a dataset—a matrix whose rows each correspond to a subject (say a person) and each column corresponds to a feature (*e.g.*, clinical measure, or the expression value of a gene), and with a final column providing the outcome (*e.g.*, a binary outcome distinguishing case versus control); see Fig 1.

These "biomarker discovery studies" (also known as "association studies") then attempt to determine which of the features (columns) differ significantly among distinct outcomes. Two standard examples of such studies are the "Genome Wide Association Studies" (GWASs), over a set of SNPs [6]; and the "Gene Signature Studies", over gene expression values [7]. Typically, this involves first computing a representative statistic for each feature (*e.g.*, for continuous entries, running a t-test based on the mean and variance of the case versus control), and then declaring a feature to be a biomarker if the resulting MCC-corrected *p*-value is below 0.05 [8], where we use "MCC" (Multiple Comparison Corrections) as a generic term, which includes both False Discovery Rate (FDR) correction, and Family-Wise Error (FWE) Correction. (Section B.I in S1 Appendix discusses some of the subtleties here, especially with respect to features).

In some situations, the researchers then validate these biomarkers using a biological or medical process (*e.g.*, based on knock-out or amplification studies [9, 10]). Other studies validate the proposed biomarkers based on existing biological knowledge. A third class of projects instead use the potential biomarkers to create a computational model–perhaps to learn a classifier [11–13]–and then measure that down-stream model (perhaps based on its accuracy on a held-out set) and declare the biomarkers to be useful if that model scores well.

A great many papers, however, simply publish the list of purported biomarkers without providing validation for this set; see Section B.2 in S1 Appendix. This paper focuses specifically on this case. We address this limitation by providing a falsifiable (statistical) claim about such sets of biomarkers, which suggests a way to validate the proposed biomarker sets.

While some biomarkers are causally related to the associated outcome, this can be difficult to establish (often requiring instrumented studies [15]); but fortunately, in many situations, it may be sufficient for the features to be *correlated* with the phenotype. Here, an ideal biomarker discovery process would identify all-and-only the features that are *consistently* correlated with the associated disease, in that its presence (or absence or . . .) alone supports that disease. This argues that a proposed biomarker is good if it was *reproducible—i.e.*, that the biomarkers found in one study, would appear in many (ideally, all) future studies that explore this disease.

This has motivated the use of independent test sets to check the validity of the earlier findings. Unfortunately, many papers report this is not the case–*i.e.*, that relatively few biomarkers appear across multiple studies. For example, while the breast cancer studies by van't Veer *et al.* [2] (resp., Wang *et al.* [3]) reported signatures with 70 (resp., 76) genes, these two sets had only 3 genes in common. Ein-Dor *et al.* [16] notes this in another situation: "Only 17 genes appeared in both the list of 456 genes of Sorlie *et al.* [17] and the 231 genes of van't Veer *et al.* [2]; merely 2 genes were shared between the sets of Sorlie *et al.* and Ramaswamy *et al.* [18]. Such disparity is not limited to breast cancer but characterizes other human disease datasets (Alizadeh *et al.* [19]) such as schizophrenia (Miklos and Maleszka [20])". Many others [21–23] report similar findings. Indeed, Begley and Ellis [24] report that only 6 of 53 published findings in cancer biology could be confirmed; which Wen *et al.* [25] notes is "a rate approaching an alarmingly low 10% of reproducibility". Moreover, a 2016 Nature survey [26], of over 1500 scientists, found that 70% of researchers have tried but failed to reproduce another scientist's experiments, and 52% thought there was a significant 'crisis' of reproducibility.

There are many possible reasons for this.

1. Each study should consider the *same well-defined "distribution" over instances—e.g.*, over the same distribution of ages and genders, etc. If the study attempts to distinguish case from control, then the two sub-populations should only differ in a single characteristic. Unfortunately, matching cases and controls over all possible features is often not achievable.

2. A second issue is with the precise notion of what "*reproducible*" means. Is it a property of a *specific biomarker*, or of *a set of biomarkers*? There is no clear choice for an optimal objective measure. This is especially problematic when dealing with multi-factorial diseases, where the outcomes correspond to a disjunction over many sub-diseases [27]. See also Section B.3 in S1 Appendix.

3. A final important issue is the impact of *sample size*. Many studies have a relatively low number of subjects, which increases the probability of finding both false negatives and false positives.

Our analysis assumes that the researchers have addressed (1) by running carefully designed, well-specified studies. Further, we also assume that there is no uncertainty in the outcomes with respect to its clinical or biological definition. We will provide a precise measure of reproducibility (2), as well as some specific implementations, and show empirically how this measurement varies with sample size (3).

In this paper, we assume that biomarkers are stand-alone features. Note each feature could be a pre-defined combination of single features (*e.g.*, the average expression values of the genes associated with a pre-defined signalling pathway—see gene enrichment [28]) or networks of genes associated with high loadings of principal component and univariate Pearson correlation values (see PC-corr [29]); but we are not considering *learning* combinations. By a *Biomarker Discovery* process BD(·), we mean a function that takes as input a labeled data matrix D of *n* subjects over a set of *r* features, each labeled with its outcome, and identifies a subset of proposed biomarkers; see Fig 1. We will formally define a *Reproducibility Score*, RS(*D*, BD), to quantify the "reproducibility" of the set of proposed biomarkers produced by the biomarker discovery process: *viz.*,

$$\begin{aligned}&\text{the average Jaccard score between these proposed biomarkers } \mathrm{BD}(D),\\&\text{and those produced by running the same BD process}\\&\text{over another comparable dataset drawn from the same distribution}\end{aligned}\tag{1}$$

where two datasets are comparable if they have the same number of subjects from each outcome. (Section B.3 in S1 Appendix discusses some subtle issues related to "reproducibility".) In order to estimate the reproducibility score in practice, we construct two approximations: an overbound and an underbound. We then provide empirical tests over many datasets, with a focus on *t*-tests as the main biomarker discovery process. We provide many examples for both microarray data (with continuous values) and SNP data (with discrete values) to provide practical evidence for the effectiveness of these approximations. Researchers can use this framework to estimate the reproducibility of the potential results of their biomarker discover study. A low reproducibility score suggests that these biomarkers may not be accurate, potentially because the dataset used is too small, the dataset is too heterogenous, or the biomarker discovery algorithm is not suitable for the dataset. To help users evaluate the quality of their proposed biomarker sets, we have also produced a publicly available website, https://biomarker.shinyapps.io/BiomarkerReprod/, that, given a labeled dataset, computes these estimates of the Reproducibility Score, with respect to any of a variety of biomarker discovery algorithms.

Outline: Section 2 formally defines the Reproducibility Score (RS) and describes the challenges of estimating this measure. We then define two approximations for RS: an overbound and an underbound. It also describes some of the standard biomarker discovery algorithms. Section 3 describes extensive empirical studies over many datasets—microarray and mRNAseq data (with continuous values) and SNP data (with discrete values), focusing on a standard Biomarker Discovery process BD(·), to confirm the effectiveness of these approximations. Section 4 summarizes some future work and the contributions of this paper. In the Supplementary Information, Section B in S1 Appendix discusses various notes: a glossary of the various technical terms used, how Biomarker Discovery differs from standard (supervised) Machine Learning, different notions of "reproducibility", how a combination of a pair of features might be important for a predictive task, even if neither, by itself, is important (towards explaining why it can be so difficult to find biomarkers); etc. Section C in S1 Appendix presents results from other empirical studies, which explore how the RS varies with the type of MCC correction used (including "none"), the p-value threshold, the size of the dataset and the number of iterations of the approximation algorithms. Finally, the approximations we present are motivated by two heuristics (Heuristics 7 and 9). Section D in S1 Appendix presents arguments that motivate these heuristics, and also provide additional empirical evidence that support them.

We close this section by motivating the need for an objective measure for evaluating the quality of a set of biomarkers (Subsection 1.1), then overviewing some earlier studies that discuss the issue of reproducibility in biomarker discovery and/or provide approaches that could be beneficial when dealing with such problems (Subsection 1.2).

1.1 Motivation for evaluating biomarker sets

To motivate the need for evaluating association studies, consider first *predictive studies*, which use a labeled dataset, like the one shown at the top of Fig 1, to produce a predictive model (perhaps a decision tree, or a linear classifier) that can be used to classify future subjects–here, into one of the two classes: Y or N. In addition to the learned classifier, the researchers will also compute *a meaningful estimate of its quality–i.e.*, of the accuracy (or AUROC, Kappa Score, etc.) of this classifier on an independent hold-out set [30], or the results of *k*-fold cross-validation over the training sample.

By contrast, many association studies report only a set of purported biomarkers, but provide no falsifiable claim about the accuracy of these biomarkers. Many meta-reviews claim that a set of biomarkers is problematic if they are not reproduced in subsequent studies [16, 31–33]. Given that biomarkers should be reproducible, we propose evaluating a biomarker set based on its reproducibility score. An accurate estimate of this score can help in at least the following three ways:

1. Researchers can compare various different "comparable" biomarker discovery algorithms to see which produces the biomarker set that is most reproducible. Here, "comparable" corresponds to the standard practice of only considering discovery tools that impose the same criterion, such as the same *p*-value, or only considering features that exhibit the same minimum fold-change. This type of analysis may help to determine errors within the biomarker discovery process.
   Moreover, we will see that MCC-correction, while useful in removing false-positives, can be detrimental to the goal of producing reproducible biomarkers; similarly, there is no reason to insist on $p < 0.05$ for the statistic test used.

2. A low reproducibility score suggests that few of the proposed biomarkers will be found in another dataset, and highlights the potential that these proposed biomarkers may not be accurate. This could motivate researchers to consider a dataset that is larger, to focus on a more homogenous population, or perhaps consider another biomarker discovery technique.

3. Finally, there are many meta-reviews [21, 34, 35] that note the lack of repeatability in many biomarker discovery papers, and question whether the techniques used are to blame. One way to address this concern is to require that each published paper include both the purported set of biomarkers, and also an estimate of its reproducibility score. The same way a prediction study's "5-fold cross validation" accuracy tells the reader how accurate the classification model should be on new data, this biomarker-discovery reproducibility score will inform the reader whether to expect another study, on a similar dataset, will find many of the same biomarkers. Note that we should view the reproducibility score as necessary for considering a proposed model, but not sufficient–*i.e.*, it might rule-out a proposed discovery model, but should not be enough to rule-in a model.

For these reasons, we provide an easy-to-use, publicly available webapp https://biomarker.shinyapps.io/BiomarkerReprod/ that anyone can use to produce meaningful estimates of the reproducibility of a set of biomarkers. (The underlying code is also available, from https://github.com/amirfrz/BMDA).

1.2 Related work

There have been many pairs of studies that have each produced biomarkers for the same disease or condition, but found little or no overlap between the two lists of purported biomarkers. Many papers have discussed this issue–some describing this problem in general [16, 33, 35], and others exploring specific examples [8, 32]. These papers suggest different reasons for the problem, such as the heterogeneous biological variations in some datasets [16, 33] or problems in the methods used that may lead to non-reproducible results [36, 37].

In particular, Zhang *et al.* [33] challenge the claim that the non-reproducibility problem in microarray studies is due to poor quality of microarray technology, by showing that inconsistencies occur even between technical replicates of the same dataset. They also show that heterogeneity in cancer pathology would further reduce reproducibility.

Ein-Dor *et al.* [16] also show the inconsistencies between the results of subsamples of a single dataset, demonstrating that the set of (gene) biomarkers discovered is not unique. They explain that there are many genes correlated with the group outcomes, but the empirical correlations change for different (sub)samples of instances. These two papers motivate our need for tools that can effectively estimate the reproducibility–such as the ones presented here.

Several projects [35–37] have attempted to formally analyse this problem. Ein-Dor *et al.* [35] describe a method, Probably Approximately Correct (PAC) sorting, that estimates the minimum number of instances needed for a desired level of reproducibility. As an example, this worst-case analysis proves that, to guarantee a 50% overlap between different gene lists for breast cancer, each dataset needs to include at least several thousand patients. This suggests poor repeatability results when using small sample sizes, which is consistent with our results for datasets with smaller sample sizes; see Subsection 3, especially Fig 4.

The goal of the MicroArray Quality Control (MAQC) project [36] was to address the problems and uncertainties about the microarray technology that were caused by the observation that different studies (of the same phenotype) often found very different biomarkers.

They suggest that the common approach of using just t-test *p*-values (especially with stringent *p*-values) can lead to poor reproducibility, which motivated them to consider methods like fold-change ranking with a non-stringent *p* cutoff, which they demonstrate leads to more reproducible gene sets. In a follow-up, Guo *et al.* [37] found similar results by using the same procedures for another dataset. However, Klebanov *et al.* [38] later show that these MAQC project results do not prove that using t-tests is necessarily unsuitable–*i.e.*, just because another method (here fold-change) can generate more reproducible results, does not mean that it is performing better; as an extreme, the algorithm that declares every gene is a biomarker (think $p = 1.0$), is completely reproducible. They demonstrate these points by using a set of simulation studies (where they know the "true biomarkers"), and use either t-test or fold-change to propose potential biomarkers. These studies found that the t-test approach performed much better than the fold-change, in terms of recall (sensitivity). These results motivated us to use the t-test approach (rather than fold-change) as our main BD algorithm–which we use for all of our empirical experiments.

Our approximation algorithms use a type of re-sampling to bound reproducibility. Below we summarize several other studies that similarly deal with re-sampling and biomarker discovery. Some studies provide ways to better estimate the true statistical significance, but do not provide a framework for evaluating empirical reproducibility—*e.g.*, Gagno *et al.* [39] used bootstrap resampling to estimate 95% confidence intervals and *p*-values for an internal assessment of their findings (related to breast cancer survival), Chitpin *et al.* [40] proposed a resampling-based method to better estimate the false discovery rate in chromatin immunoprecipitation experiments, and Pavelka *et al.* [41] proposed a resampling-based hypothesis testing algorithm that provides a control of the false positive rate for identification of differentially expressed genes. Furthermore, Alshawaqfeh *et al.* [42] and Zhao and Li [43] suggested methods for consistent biomarker detection in high-throughput datasets where evaluation was based on common biomarkers among the two resampled sets–*i.e.*, they are considering the false positives but not false negatives. (By contrast, our use of Jaccard score involves both.) Other studies had different goals–*e.g.*, Ma *et al.* [44] used resampling in a permutation test, to evaluate the predictive power of the identified gene set based on the accuracy of downstream classification task; recall however that our goal is to evaluate the reproducibility of the biomarkers directly (not downstream). Filosi *et al.* [45] propose methods for evaluating the stability of reconstructed biological networks in terms of inference variability due to data subsampling; we however are focusing on the reproducibility of the individual biomarkers. Hua *et al.* [46] conducted a simulation study to compare the *ranking performance* of several gene set enrichment methods; by contrast, our approach considers the SET of biomarkers, not the ranking, and is over several real-world datasets (not just simulated ones). Note that none of these used re-sampling techniques to bound the expected replicability of the set of biomarkers found by some discovery algorithm, nor to demonstrate the validity of those bounds.

2 Materials and methods

2.1 Formal description

As illustrated in Fig 1, a "Biomarker Discovery" algorithm, BD($\cdot$), takes as input a dataset D of n subjects, each described by r features $F = \{f_1, \ldots, f_r\}$ and labeled with a binary outcome, and returns a subset $F' \subset F$ of purported biomarkers.

Typically, each $f \in F'$ differs in some significant way between each class. To be more precise, let x_i^j be the value of the i^{th} feature of the j^{th} subject, and $\ell^{(j)}$ be the outcome of the j^{th} subject (which is either + or -). Then a class difference means that the set $\{x_i^j \mid \ell^{(j)} = +\}$ of values of the i^{th} feature of the diseased individuals is significantly different from the values of that feature over the healthy individuals, $\{x_i^j \mid \ell^{(j)} = -\}$. For simplicity, we will assume that these $\{x_i^j\}_{i,j}$ values are either all continuous (such as height, or the expression value of a gene), or all discrete (such as gender, or the genotype of a SNP). Subsection 2.2 below will describe several such biomarker discovery algorithms.

As noted in Eq 1, the *Reproducibility Score* RS(D, BD) quantifies the "reproducibility" of the set of proposed biomarkers BD(D) corresponding to running the biomarker discovery algorithm over the labeled dataset D. Here, we assume that the values of each feature $x_\cdot^j$, for each outcome c, are generated independently from a fixed distribution (*i.e.*, "i.i.d.")

$$p_{i,c}(v) \quad = \quad P(\, x_i^j = v \mid \ell^{(j)} = c\,).$$

Here and in general, we use $P(\cdot)$ to refer to either a probability density for continuous variables, or a probability mass for discrete variables. Note these are just the marginal distributions: we do not assume that the various features are independent from one another.–*i.e.*, this does not necessarily correspond to Naive Bayes [30]. We will view $\vec{p}(\cdot) = [p_{i,c}(\cdot)]_{i,c}$ as the matrix of these $r \times 2$ different distributions, and let $\vec{p}^{[n_+,n_-]}(\cdot)$ be the distribution for sampling $n = n_+ + n_-$ instances independently from this distribution, where $n_+ \in \mathbb{Z}^+$ instances are drawn from the distribution $[p_{1,+}(\cdot), \ldots, p_{r,+}(\cdot)]$ associated with positive outcomes, and similarly $n_- \in \mathbb{Z}^+$ instances from the distribution $[p_{1,-}(\cdot), \ldots, p_{r,-}(\cdot)]$ associated with negative outcomes. Then for datasets $D', D'' \sim \vec{p}^{[n_+,n_-]}(\cdot)$ sampled independently, we define

$$\text{RS}^*(\,\vec{p}(\cdot),\ [n_+, n_-],\ \text{BD}(\cdot)) \quad = \quad \mathbb{E}[\,J(\text{BD}(D'),\ \text{BD}(D''))\,] \tag{2}$$

where the Jaccard score of two sets

$$J(A,\ B\,) \quad = \quad \frac{|\,A\ \cap\ B\,|}{|\,A\ \cup\ B\,|} \tag{3}$$

is the ratio of the intersection to the union of these sets—hence $J(A, B)$ ranges from 0 to 1, and is 1 if and only if $A = B \neq \{\}$, and is 0 if and only if these sets are disjoint. (We define this to be 0 if $A = B = \{\}$.) Note that the Jaccard score is only one possible measure to evaluate the degree of overlap of gene signatures; Section C.4 in S1 Appendix discusses a slightly different measure that is sometimes used to measure the reproducibility of a set of biomarkers. See also Shi *et al.* [47] for a comprehensive overview of these measures.

Of course, we do not know $\vec{p}(\cdot)$, and so we use an empirical distribution $\widehat{p_D}(\cdot)$, determined based on context from the dataset D, to produce the approximation

$$\mathrm{RS}(\,D,\ \mathrm{BD}) \quad = \quad \mathrm{RS}^*(\,\widehat{p_D}(\cdot),\ \langle D\rangle,\ \mathrm{BD}(\cdot)) \tag{4}$$

that estimates the reproducibility of the biomarker set BD(D), where the notation $\langle D\rangle = [|D^+|, |D^-|]$ refers to the pair of sizes of the positive and negative subjects in D—corresponding to $[n_+, n_-]$. Note that this reproducibility score deals with the *sets* of biomarkers that are produced by the BD($\cdot$) function, and not any specific biomarker.

Of course, Eq 4 suggests the obvious bootstrap sampling algorithm [48]. Empirically, however, we found that it did not perform well (see Section C.1 in S1 Appendix)–motivating the algorithms described in Subsection 2.3.

2.2 Biomarker discovery algorithms: BD($\cdot$)

We now discuss various approximation algorithms for biomarker discovery. As our goal is to illustrate the reproducibility issues with respect to the standard approach, we focus on that standard approach: where the biomarker discovery is based on independent two-sample *t*-tests, perhaps with some multiple comparison correction. This use of t-tests implicitly assumes that the feature values, for each outcome, are normally distributed. However, it is easy to adapt these algorithms to use other statistical tests, that do not make this distributional assumption.

See also Limitations in Section 4.

Recall that we are considering two types of datasets, depending on whether its feature values (the x_i^j mentioned above) are continuous or discrete. However, for datasets with categorical values—SNPs in our analysis–we use a simple preprocessing step, which precedes all the BD($\cdot$) algorithms described here, to convert each categorical value to a real number: here converting each SNP feature, which ranges over the values { AA, Ab, bb }, to the real-values { 0, 1, 2 }, corresponding to the number of minor alleles ("b") in the genotype. This allows us to view each such dataset as one with continuous values.

We assume that the real values of each feature, for each outcome, follows a normal distribution, which might be different for the different outcomes, and so we use an independent two-sample *t*-test for all of our empirical experiments. Recall that the test statistic is given by

$$\mathrm{t} \quad = \quad \frac{\bar{X}_+ \ - \ \bar{X}_-}{\bar{s}_p \cdot \sqrt{\dfrac{1}{n_+} \ + \ \dfrac{1}{n_-}}} \qquad\qquad \bar{s}_p \quad = \quad \sqrt{\frac{(n_+ - 1)\,\bar{s}_+^2 \ + \ (n_- - 1)\,\bar{s}_-^2}{n_+ \ + \ n_- - 2}}. \tag{5}$$

where n_+ and n_- are the number of instances with positive and negative outcomes, respectively, and with empirical means $\bar{X}_-$ and $\bar{X}_+$ and empirical variances $\bar{s}_+^2$ and $\bar{s}_-^2$.

Note the biomarker discovery process essentially performs a single statistical test for each feature. As the number of features is often large–often tens-of-thousands, or more–many projects sought ways to reduce the chance of false discoveries; a standard way to do this is through some MCC method. We therefore consider biomarker discovery algorithms described as $\mathrm{BD}_{t,\, p,\, \chi}(D)$, where the t in the subscript refers to the 2-sided t-test, the p for the p-value used, and χ to the MCC method. Our canonical example is $\mathrm{BD}_{t,\, 0.05,\, BH}(D)$, with $p = 0.05$, and $\chi = \mathrm{BH}$ to the Benjamini/Hochberg correction [49]. This notation makes it easy to consider many variants –- *e.g.*, adjusting the p-value used for the statistical test, whether it is applying another multiple testing correction, or none, etc. See Section C.2 in S1 Appendix for more details.

2.3 Algorithms that approximate the reproducibility score

As we have the dataset D with $n \approx [n_+, n_-]$ labeled instances, we can directly compute $\mathrm{BD}(D)$. To compute $\mathrm{RS}(D, \mathrm{BD}(\cdot))$, however, we also need to produce one (or more) similar datasets D', each with $[n_+, n_-]$ subjects drawn from the same (implicit) distribution $\vec{p}(\cdot)$ that generated D (with the same number of positive and negative instances), but which is presumably disjoint from D. While we do not have such D''s, and so cannot directly compute the Reproducibility Score, we show below how to compute an overbound and an underbound of $\mathrm{RS}(D, \mathrm{BD}(\cdot))$.

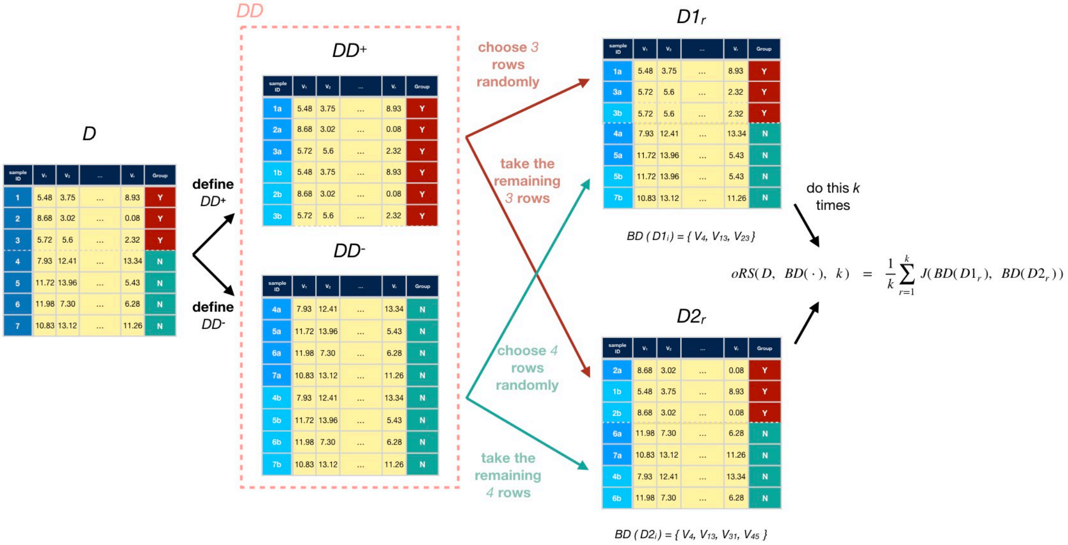
2.3.1 Overbound: oRS. The $\mathrm{oRS}(D, \mathrm{BD}(\cdot), k)$ procedure produces (an estimate of) an overbound for $\mathrm{RS}(D, \mathrm{BD}(\cdot))$, by making it *easier* for a feature to be selected to be in both purported biomarker sets. In this algorithm, D is our given fixed dataset, $\mathrm{BD}(\cdot)$ is the given biomarker discovery algorithm, and k is a parameter to determine the number of trials when computing the overbound. (Here, and below, see the Glossary in Section A in S1 Appendix for a summary of the algorithms and their arguments.) The oRS algorithm first defines a size-$2n$ dataset DD that contains two copies of each subject in D, of course with the same outcome both times. It then randomly partitions this DD into two disjoint size-n datasets D_1 and D_2, balanced by outcome. Here, each partition is with respect to the *list* of elements, so it will include duplicates. To insure that the resulting datasets are balanced, oRS first split DD into DD^+ and DD^-, where DD^+ are the cases and DD^- the controls. It then forms D_1^+ by randomly drawing 1/2 of DD^+, and D_1^- by randomly drawing 1/2 of DD^-, then merges $D_1 = D_1^+ \cup D_1^-$; see Fig 2. The dataset D_2 is then formed from the remaining subjects of DD not included in D_1. oRS then runs $\mathrm{BD}(\cdot)$ on the datasets D_1 (resp., D_2) to produce two sets of biomarkers, and computes the Jaccard score for this pair of biomarker sets: $J(\mathrm{BD}(D_1), \mathrm{BD}(D_2))$. It then repeats this double-split-BD-Jaccard process k times, then returns the average of these k values:

$$\mathrm{oRS}(\, D,\, \mathrm{BD}(\cdot),\, k) \quad = \quad \frac{1}{k}\sum_{i=1}^{k} J(\mathrm{BD}(D1_r),\ \mathrm{BD}(D2_r)\ ) \tag{6}$$

where each dataset pair $[D1_r, D2_r]$ is created independently using the above procedure.

For each r, as we expect $D1_r$ to overlap with $D2_r$, it is relatively likely that any $D1_r$-biomarker will also be a $D2_r$-biomarker (more likely than if $D1_r$ was disjoint from $D2_r$), which means we expect the associated Jaccard score to be higher. This follows from the heuristic that, as two datasets have more common elements, we expect the number of biomarkers common to two datasets, to increase–*i.e.*,

$$\begin{array}{cccc} \text{if} & A_1, A_2, B_1, B_2 \sim \vec{p}^{[n_+, n_-]}(\cdot) & \text{and} & \\ & |A_1 \cap A_2| & \text{is larger than} & |B_1 \cap B_2|, \\ \text{then we expect} & J(\mathrm{BD}(A_1), \mathrm{BD}(A2)) & \text{will be larger than} & J(\mathrm{BD}(B_1), \mathrm{BD}(B2)) \end{array} \tag{7}$$


D
define
DD+
define
DD-
DD
DD+
DD-
choose 3
rows
randomly
take the
remaining
3 rows
choose 4
rows
randomly
take the
remaining
4 rows
D1r
BD (D1i) = { V4, V13, V23 }
D2r
BD (D2i) = { V4, V13, V31, V45 }
do this k
times
$oRS(D,\ BD(\cdot),\ k)\ =\ \frac{1}{k}\sum_{r=1}^{k} J(BD(D1_r),\ BD(D2_r))$

sample ID	V_1	V_2	...	V_r	Group
1	5.48	3.75	...	8.93	Y
2	8.68	3.02	...	0.08	Y
3	5.72	5.6	...	2.32	Y
4	7.93	12.41	...	13.34	N
5	11.72	13.96	...	5.43	N
6	11.98	7.30	...	6.28	N
7	10.83	13.12	...	11.26	N

sample ID	V_1	V_2	...	V_r	Group
1a	5.48	3.75	...	8.93	Y
2a	8.68	3.02	...	0.08	Y
3a	5.72	5.6	...	2.32	Y
1b	5.48	3.75	...	8.93	Y
2b	8.68	3.02	...	0.08	Y
3b	5.72	5.6	...	2.32	Y

sample ID	V_1	V_2	...	V_r	Group
4a	7.93	12.41	...	13.34	N
5a	11.72	13.96	...	5.43	N
6a	11.98	7.30	...	6.28	N
7a	10.83	13.12	...	11.26	N
4b	7.93	12.41	...	13.34	N
5b	11.72	13.96	...	5.43	N
6b	11.98	7.30	...	6.28	N
7b	10.83	13.12	...	11.26	N

sample ID	V_1	V_2	...	V_r	Group
1a	5.48	3.75	...	8.93	Y
3a	5.72	5.6	...	2.32	Y
3b	5.72	5.6	...	2.32	Y
4a	7.93	12.41	...	13.34	N
5a	11.72	13.96	...	5.43	N
5b	11.72	13.96	...	5.43	N
7b	10.83	13.12	...	11.26	N

sample ID	V_1	V_2	...	V_r	Group
2a	8.68	3.02	...	0.08	Y
1b	5.48	3.75	...	8.93	Y
2b	8.68	3.02	...	0.08	Y
6a	11.98	7.30	...	6.28	N
7a	10.83	13.12	...	11.26	N
4b	7.93	12.41	...	13.34	N
6b	11.98	7.30	...	6.28	N

Fig 2. Computing oRS. Diagram showing oRS's process of generating pairs of subsets for a dataset *D* (*k* times), then using those to compute oRS(D, BD(·), k).

https://doi.org/10.1371/journal.pone.0252697.g002

ceteris paribus. Here, we view each of $\{A_1, A_2, B_1, B_2\}$ as a set of $n = n_+ + n_- r$-dimensional instances. Note this relationship is simply a heuristic to motivate the algorithm–one that we expect to hold in practice. Section D in S1 Appendix provides some basic arguments, and empirical evidence, to support this claim.

We close with three quick observations:

1. **Expected Overlap:** We expect 50% of the instances to be duplicated in any given pair of datasets. See Lemma 2 in Section D.4 in S1 Appendix.
2. **Relation to Bootstrap Samples:** Here, each subject occurs exactly twice in each pair of datasets $D1_r$ and $D2_{,r}$. If we instead used bootstrap sampling (called bRS below), we expect many subjects would occur more often in the pair of datasets. (See Lemma 3 in Section D.4 in S1 Appendix.) Given Heuristic 7, this means we expect bRS's Jaccard score here to be higher than for oRS's; as oRS is already an overbound for the true score (RS), this means bRS would be a worse bound, as bRS ≥ oRS ≥ RS. This is why we use our "doubling approach" oRS rather than bootstrap sampling bRS, as oRS produces values that are smaller, but still remains an overbound, as desired. See Figs 7 and 8 and Section C.1 in S1 Appendix.
3. **Relation to RS*:** Our oRS approach is clearly related to RS* (Eq 2), as both compute the average Jaccard score of pairs of size-n datasets sampled from a distribution. They differ as (a) RS* uses the true distribution $\vec{p}(\cdot)$, while oRS uses only the estimate $\hat{p}_D(\cdot)$, (b) RS* is the *true* average while oRS is just the *empirical* average over k trials, and (c) RS* will draw *independent* datasets, but the oRS datasets will overlap.

2.3.2 Underbound: uRS. The uRS(D, BD, k) procedure produces (an estimate of) an underbound for RS(D, BD(·)) by making it *harder* for a feature to be selected to be in both purported biomarker sets. First, observe that as $[n_+, n_-]$ increases (keeping the n_+-to-n_- ratio fixed, as we consider changing the size of the dataset), we expect the statistical estimates to be more accurate, and in particular, statistical tests for differences between the two classes will be correct more often. Hence, a statistical test will better identify the "true" biomarkers F^* from a size-n subset $D^{(n)}$, versus from a size-$n/2$ subset $D^{(n/2)}$. Now consider two size-n datasets $D_1^{(n)}$ and $D_2^{(n)}$, and also two size-$n/2$ datasets $E_1^{(n/2)}$ and $E_2^{(n/2)}$. As $\mathrm{BD}(D_1^{(n)})$ and $\mathrm{BD}(D_2^{(n)})$ are each closer to F^* than $\mathrm{BD}(E_1^{(n/2)})$ and $\mathrm{BD}(E_2^{(n/2)})$, we expect $\mathrm{BD}(D_1^{(n)})$ and $\mathrm{BD}(D_2^{(n)})$ to be closer to each other, than $\mathrm{BD}(E_1^{(n/2)})$ and $\mathrm{BD}(E_2^{(n/2)})$, which means we expect that

$$J(\mathrm{BD}(D1^{(n)}),\ \mathrm{BD}(D2^{(n)})\ ) \quad \geq \quad J(\mathrm{BD}(E1^{(n/2)}),\ \mathrm{BD}(E2^{(n/2)})\ ).$$

In general, given that

$$\widehat{E}^{(k)}[\ J(\mathrm{BD}(\mathrm{D1}^{(s)}),\ \mathrm{BD}(\mathrm{D2}^{(s)})\ )\ ] \quad \approx \quad \mathrm{RS}^*(\vec{p}(\cdot),\ s,\ \ \mathrm{BD}(\cdot)) \tag{8}$$

(where $\widehat{E}^{(k)}[\cdot]$ is the empirical average over k samples), this argues that the RS^* score should increase with the size s of the dataset–which suggests that

$$\begin{array}{llll} \text{if} & A_1, A_2, \sim \vec{p}^{sA}(\cdot), B_1, B_2, \sim \vec{p}^{sB}(\cdot) & \text{and} & \\ & sA = |A_1| = |A_2| & \text{is larger than} & sB = |B_1| = |B_2|, \\ \text{then we expect} & J(\mathrm{BD}(A_1), \mathrm{BD}(A2)) & \text{will be larger than} & J(\mathrm{BD}(B_1), \mathrm{BD}(B2)) \end{array}$$

Fig 4(a) presents empirical evidence, over 5 datasets, supporting this claim —showing that the Jaccard score increases as we increase the size s of the datasets. Section D in S1 Appendix provides some arguments, and additional empirical evidence (over hundreds of simulations), that further support this heuristic.

This motivates our underbound algorithm uRS(D, BD, k), which first partitions D into two disjoint size-$n/2$ subsets, E_1 and E_2, with balanced outcomes. It then computes $J(\mathrm{BD}(E_1), \mathrm{BD}(E_2))$ which, assuming Heuristic 9, is an underbound in expectation for RS(D, BD(·)). uRS does this partitioning k times, producing k different dataset pairs $\{[E_{1,r}, E_{2,r}]\}_{r = 1, \ldots, k}$, and returning the average Jaccard score, i.e.

$$\mathrm{uRS}(D,\ \mathrm{BD}(\cdot),\ k\ ) \quad = \quad \frac{1}{k}\sum_{r=1}^{k} J(\mathrm{BD}(E_{1,r}),\ \mathrm{BD}(E_{2,r})\ ). \tag{10}$$

See Fig 3.

2.4 Empirical study over various datasets

There are now many publicly-available datasets that have been used in association studies. Here, we use them to . . .

U1: Better understand what Jaccard scores are typical, for the standard BD(·) algorithms;

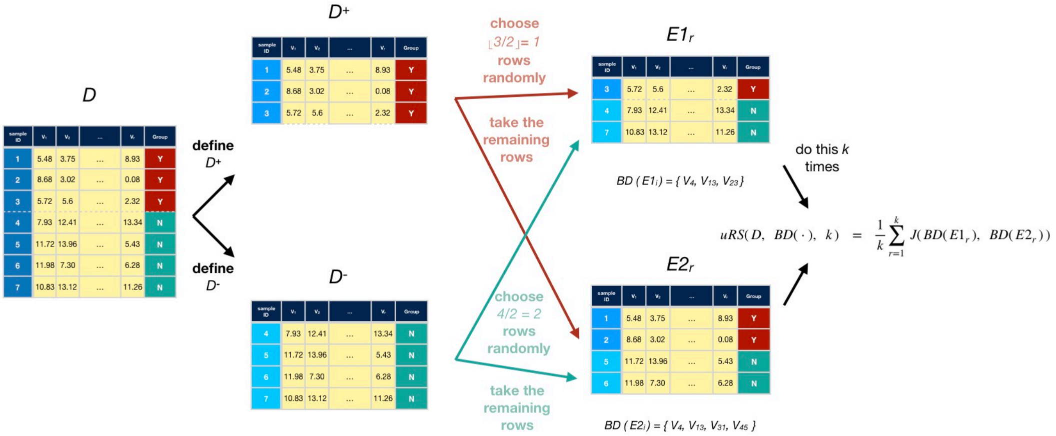
U2: Determine whether our predictions match the results of earlier meta-analyses; and

U3: Determine if our approximations are meaningful—*i.e.*, if (for reasonable values of k):

$$\mathrm{uRS}(\ D,\ \mathrm{BD},\ k) \quad \leq \quad \mathrm{RS}(\ D,\ \mathrm{BD}) \tag{11}$$

$$\mathrm{oRS}(\ D,\ \mathrm{BD},\ k) \quad \geq \quad \mathrm{RS}(\ D,\ \mathrm{BD}) \tag{12}$$

The next section will explicitly discuss all three issues. Of course, given only a single dataset D of size-n, we cannot compute, nor even estimate, the true value of RS(D, BD(·)). However,



D

sample ID	V_1	V_2	...	V_r	Group
1	5.48	3.75	...	8.93	Y
2	8.68	3.02	...	0.08	Y
3	5.72	5.6	...	2.32	Y
4	7.93	12.41	...	13.34	N
5	11.72	13.96	...	5.43	N
6	11.98	7.30	...	6.28	N
7	10.83	13.12	...	11.26	N

D+

sample ID	V_1	V_2	...	V_r	Group
1	5.48	3.75	...	8.93	Y
2	8.68	3.02	...	0.08	Y
3	5.72	5.6	...	2.32	Y

D-

sample ID	V_1	V_2	...	V_r	Group
4	7.93	12.41	...	13.34	N
5	11.72	13.96	...	5.43	N
6	11.98	7.30	...	6.28	N
7	10.83	13.12	...	11.26	N

E1r

sample ID	V_1	V_2	...	V_r	Group
3	5.72	5.6	...	2.32	Y
4	7.93	12.41	...	13.34	N
7	10.83	13.12	...	11.26	N

$BD\,(E1_i) = \{V_4, V_{13}, V_{23}\}$

E2r

sample ID	V_1	V_2	...	V_r	Group
1	5.48	3.75	...	8.93	Y
2	8.68	3.02	...	0.08	Y
5	11.72	13.96	...	5.43	N
6	11.98	7.30	...	6.28	N

$BD\,(E2_i) = \{V_4, V_{13}, V_{31}, V_{45}\}$

$$uRS(D,\ BD(\cdot),\ k) \;=\; \frac{1}{k}\sum_{r=1}^{k} J(BD(E1_r),\ BD(E2_r))$$

Fig 3. Computing uRS. Diagram showing uRS's process of generating pairs of subsets for a dataset D (k times), then using those to compute uRS(D, BD(·), k).

https://doi.org/10.1371/journal.pone.0252697.g003

we can estimate RS($D^{(n/2)}$, BD(·)), where $D^{(n/2)}$ is a size-n/2 (outcome-balanced) subset of D. In fact, uRS(D, BD(·), k) is a meaningful estimate of RS($D^{(n/2)}$, BD(·)); below we will use

$$\widehat{\mathrm{RS}}(\, D^{(n/2)},\ \mathrm{BD}(\cdot),\ k) \quad = \quad \mathrm{uRS}(\, D,\ \mathrm{BD}(\cdot),\ k) \ . \tag{13}$$

We will then compare this $\widehat{\mathrm{RS}}(\, D^{(n/2)},\ \mathrm{BD}(\cdot),\ k)$ against uRS($D^{(n/2)}$, BD, k) and oRS($D^{(n/2)}$, BD, k) and to see whether the relations of Eqs 11 and 12 both hold, with respect to various size-n/2 subsets $D^{(n/2)}$.

More generally, we can do this for any size-s subset $D^{(s)}$ of D where $s \leq n/2$. Here, we need a set of pairs of disjoint outcome-balanced subsets D', $D'' \subset D$ where $|D'| = |D''| = s$ and $D' \cap D'' = \{\}$. For a fixed dataset D, and specified number $k \in \mathbb{Z}^{>0}$, we can then plot these $\widehat{\mathrm{RS}}(\, D^{(s)},\ \mathrm{BD}(\cdot),\ k)$ values along with oRS($D^{(s)}$, BD, k) and uRS($D^{(s)}$, BD, k), as a function of s to see their behaviour; see Fig 4(b), for the Metabric dataset. Our website https://biomarker.shinyapps.io/BiomarkerReprod/ also provides this visualization.

We explored our approximations over **25 different real-world datasets, including 16 microarray datasets and 2 RNAseq datasets with continuous data, and 7 SNP datasets with**

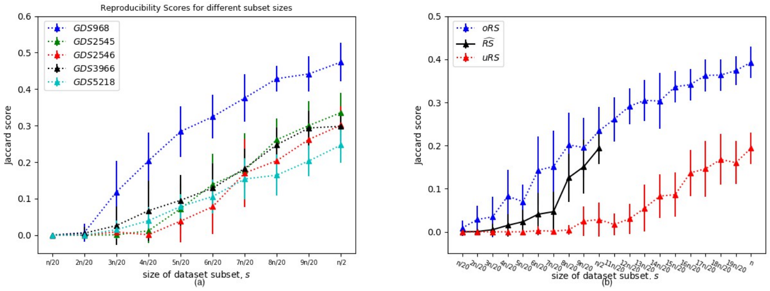


Fig 4. Reproducibility Scores for different subset sizes. (a) For each of the 5 datasets D, each point shows the average (± sd) Jaccard $\widehat{E}^{(k)}[\,J(\mathrm{BD}(D1^{(s)}),\ \mathrm{BD}(D2^{(s)})\,)\,]$ over $k = 20$ pairs [$D1^{(s)}$, $D2^{(s)}$] of disjoint size-s subsets of D. Here, n is the size of the original dataset–note this can only go to n/2–and we are using the standard $\mathrm{BD}_{t,\,0.05,\,BH}$. (b) Showing how the approximations relate to one another, and scale with the size s of the dataset. Here we are using subsets of the Metabric dataset, with n = 1654. We observed the same behavior for all datasets.

https://doi.org/10.1371/journal.pone.0252697.g004

Table 1. Results for all 25 datasets when using all the subjects. This table is sorted by sample size (#subjects)–corresponding to Fig 5. The first 18 entries are Gene Expression datasets (including the 4 "*"ed entries, from the Zou *et al.* [8] meta-study), and the final 7 are SNP datasets. Reproducibility Scores are shown in the form of mean ± standard deviation. The "(Majority %)" values are the percentage of the subjects in the dataset with the more common outcome–*e.g.*, 53% of the subjects in the GDS968 dataset are labeled "+" (for "long survival time"), and 82% of GSE7390 are labeled "–" (for "Non-Metastatic").

Name	#subjects (Majority %)	#features	#biomarkers	uRS %	oRS %
GDS968 [50]	171 (53%)	5748	2506	47.5 ± 3.95	63.25 ± 0.76
GSE7390* [14, 51]	198 (82%)	13245	18	0 ± 0	5.15 ± 2.81
GSE2034* [3]	286 (67%)	13245	277	0 ± 0	12.6 ± 4.19
GSE1456* [52]	159 (78%)	13245	443	0 ± 0	13.6 ± 6.4
GSE11121* [53]	200 (86%)	13245	492	0.09 ± 0.2	13.4 ± 5.89
GDS2546 [54, 55]	167 (54%)	12553	2965	30.8 ± 4.82	49.0 ± 0.55
GDS2545 [54, 55]	171 (53%)	12558	4291	34.0 ± 4.09	54.58 ± 0.31
GDS2547 [54, 55]	164 (54%)	12579	1810	23.7 ± 5.86	42.66 ± 0.70
KIPAN [56]	532 (81%)	18271	2782	12.3 ± 4.91	34.2 ± 4.74
BRCA [56]	552 (95%)	18320	2	0 ± 0	2.5 ± 2.1
GDS2771 [57, 58]	187 (52%)	22215	1807	0.32 ± 0.64	31.68 ± 0.47
GDS3966 [59]	83 (63%)	22274	6554	31.7 ± 4.71	53.66 ± 0.27
GDS4185 [60, 61]	67 (58%)	22283	6	0 ± 0	11.29 ± 0.83
Metabric [62]	1654 (57%)	24368	3675	18.5 ± 3.86	39.8 ± 3.56
GDS4431 [63]	146 (53%)	54613	140	0 ± 0	18.85 ± 0.34
GDS4719 [64]	19 (53%)	54675	1	0 ± 0	7.02 ± 0.71
GDS2737 [65]	37 (57%)	54675	4	0 ± 0	10.58 ± 0.5
GDS5218 [66]	110 (56%)	54675	10700	24.0 ± 5.09	46.06 ± 0.29
GSE13429 [67]	39 (79%)	262314	1267	1.66 ± 0.44	21.4 ± 3.76
GSE25103 [68]	122 (92%)	908512	325	0.27 ± 0.14	3.94 ± 2.22
GSE15097 [69]	68 (59%)	909456	108224	4.9 ± 2.9	34.2 ± 5.64
GSE15096 [69]	69 (58%)	909457	106482	5.37 ± 2.9	33.4 ± 5.28
GSE25104 [68, 70]	122 (92%)	909547	326	0.27 ± 0.14	3.94 ± 2.22
GSE15826 [71]	164 (54%)	909549	0	0 ± 0	2.5 ± 0.58
GSE18333 [72]	82 (54%)	909606	0	0 ± 0	0 ± 0

https://doi.org/10.1371/journal.pone.0252697.t001

categorical data (see Table 1). This first set includes 4 of the gene expression datasets discussed in the Zou *et al.* [8] meta-analysis—each describing metastatic versus non-metastatic breast primary cancer subjects—to see if our method is consistent with their empirical results. We also included 11 other relatively-small gene expression datasets (with 19 to 187 subjects), focusing on human studies that had a binary class outcome from the GEO repository. To explore how our tools scale with size, we also included 3 other relatively large datasets, with 532 to 1654 subjects. As these were survival datasets, we set the binary outcome based on the median survival time (removing any subject that was censored before that median time). In addition to these 4+11+3 = 18 gene expression datasets (with real-valued entries), we also include 7 SNP datasets (with 39 to 164 subjects), with discrete values, also selected from human studies with binary class outcome s. Fig 5 plots the number of features and biomarkers found, using the $\text{BD}_{t,\,0.05,\,BH}$ algorithm, for each dataset—both $D^{(n)}$ and $D^{(n/2)}$.

Results

We ran our suite of methods over the aforementioned 25 datasets, including 16 microarray datasets and 2 mRNAseq datasets, whose feature-values $\{x_i^j\}$ (recall each x_i^j is the expression value of the i-th gene for the j-th subject; we $\log_2$-transformed the values from the mRNAseq

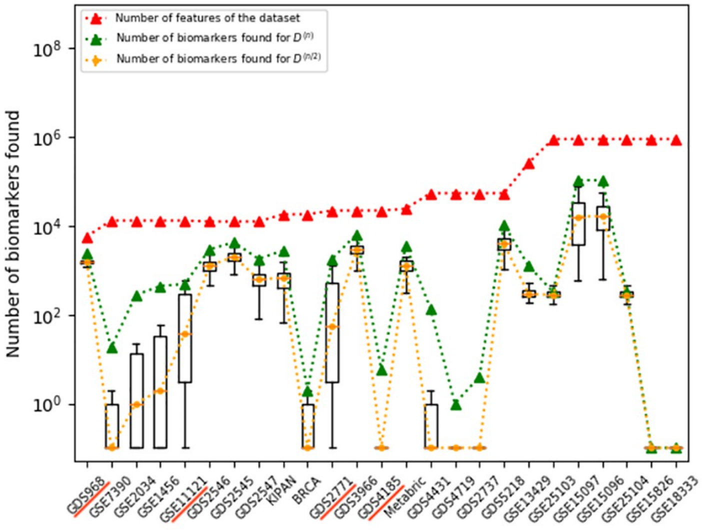


Fig 5. Number of biomarkers found for $D^{(n/2)}$ compared to $D^{(n)}$. Box+whiskers plots showing number of biomarkers found for $D^{(n/2)}$ when using $\mathrm{BD}_{t,\,0.05,\,BH}$ over $k = 20$ iterations for various dataset, compared to the number of biomarkers for $D^{(n)}$, and to the number of features in each dataset. Note the y-axis is a log-scale. (Note we first changed all "0" values to "10^{-1}".) For details, see Tables 1 and 2.

https://doi.org/10.1371/journal.pone.0252697.g005

datasets) and 7 were SNP datasets with categorical entries, *i.e.*, $x_i^j \in \{0, 1, 2\}$ is the number of minor alleles in the genotype for the i^{th} SNP for the j^{th} subject; see Table 1. Here, we use the standard $\mathrm{BD}_{t,\,0.05,\,BH}(\cdot)$ biomarker discovery algorithm; see Section 2.2.

First, to address (U1) and (U2) in Section 2.4, we analyzed the 4 datasets mentioned in the Zou *et al.* [8] meta-analysis (see the 4 "*" rows of Table 1) and computed the $\{\mathrm{uRS}(D, \mathrm{BD}_{t,\,0.05,\,BH}, 50), \mathrm{oRS}(D, \mathrm{BD}_{t,\,0.05,\,BH}, 50)\}$ values for each dataset D, as well as the actual Jaccard score for biomarkers for each pair of datasets. (We were not able to replicate the results reported for the 5th dataset from that study, and so we excluded that one dataset from our analysis.) The results, in Fig 6, show that the Jaccard score for each pair is well within the bounds computed by our approximations, for each of the datasets in that pair—that is, the results for 4 × 3 = 12 ordered-pairs of datasets are consistent with our predictions. Note that we verified that our $\mathrm{BD}_{t,\,0.05,\,BH}$ algorithm matched the original study by verifying that the PO scores (Eq 15 in Section C.4 in S1 Appendix) matched the ones that were originally published.

To address (U3), we also analyzed the other 14 continuous datasets D, and computed the oRS and uRS values using $k = 50$ repetitions; see Fig 7[left]. We see that the overbound oRS is consistently larger than the underbound uRS—*i.e.*, $\mathrm{uRS} \leq \mathrm{oRS}$—as claimed by Eqs 11 and 12. Fig 7[right] plots the corresponding values for the $D^{(n/2)}$ datasets, that use only 1/2 of the dataset, using the same $\mathrm{BD}(\cdot)$ algorithm and $k = 50$. It also plots the "true" $\widehat{\mathrm{RS}}(\,D^{(n/2)},\ \mathrm{BD}(\cdot),\ k)$ values for the datasets. Again, we see that $\mathrm{oRS} \geq \widehat{\mathrm{RS}} \geq \mathrm{uRS}$.

Finally, similar to that experiment over the 18 continuous datasets, we examined the 7 discrete datasets and produced the reproducibility scores. Fig 8[left] shows the scores for each of the 7 SNP datasets, demonstrating that $\mathrm{oRS} \geq \mathrm{uRS}$ holds for the discrete cases as well. Fig 8 [right] shows the scores for $D^{(n/2)}$ datasets, and performs the same verification. Those figures also allow us to see the Jaccard scores (U1 above) range from essentially 0 to around 0.475 for the $D^{(n/2)}$ datasets. In addition to the plots, Tables 1 and 2 present the relevant values.

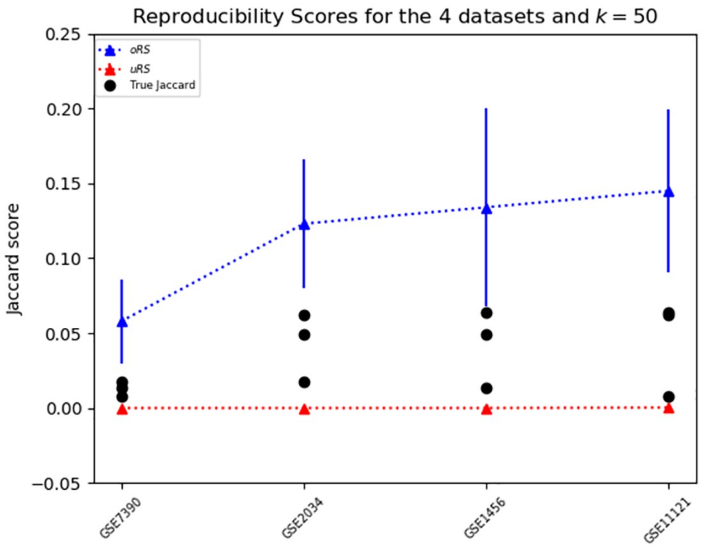


Fig 6. oRS and uRS compared to real Jaccard scores. Under-bound and over-bound for the 4 datasets from Zou *et al.* [8], as well as the true Jaccard score for each pair—3 numbers for each dataset, shown by black circles.

https://doi.org/10.1371/journal.pone.0252697.g006

Section C in S1 Appendix provides the results of many additional empirical studies, showing how the reproduciblity scores change based on which (if any) MCC method is used, the specific *p*-value used for the t-test, the number of draws *k* used by the various approximations, and the size of the dataset *n*.

4 Discussion

4.1 Recommended use

As suggested above, whenever researchers have identified a possible set of biomarkers from a dataset *D* and Biomarker Discovery Algorithm BD, we encourage them to apply our analytic technique, and where appropriate, even our specific bounding algorithms (from https://

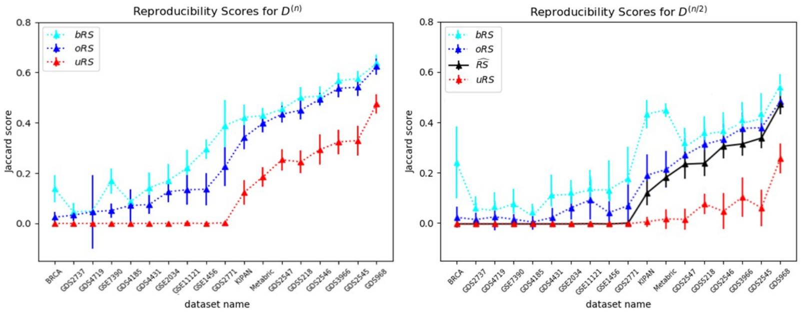


Fig 7. Empirical results for continuous datasets. Reproducibility scores (mean and standard deviation) for all 18 continuous datasets, both for complete datasets with n subjects (left) and for half-sized with $\frac{n}{2}$ subjects (right), for $k = 50$ iterations. The x-axes (for both plots) are sorted by the value of the over-bound for the $D^{(n)}$ datasets. We see, in both, that the over-bound oRS is consistently higher than the under-bound uRS. Moreover, the right plot shows that the "truth" $\widehat{\text{RS}}$ is also between uRS and oRS. (The bRS lines in the plots are based on the Bootstrap Overbound method; see Section C.1 in S1 Appendix).

https://doi.org/10.1371/journal.pone.0252697.g007

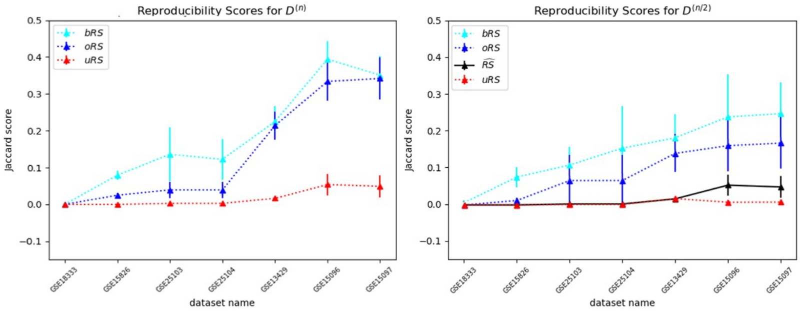


Fig 8. Empirical results for SNP datasets. Reproducibility scores (mean and standard deviation) for 7 SNP datasets, both for complete datasets with n subjects $D^{(n)}$ (left) and for half-sized with $\frac{n}{2}$ subjects $D^{(n/2)}$ (right), for 50 iterations. The x-axes (for both plots) is sorted by the value of the overbound oRS for the $D^{(n)}$ datasets. We see, in both, that the over-bound oRS is consistently higher than the under-bound uRS. Moreover, the right plot shows that the "truth" $\widehat{\text{RS}}$ is also within the range of oRS and uRS. (The bRS lines in the plots are based on the Bootstrap Overbound method; see Section C.1 in S1 Appendix).

https://doi.org/10.1371/journal.pone.0252697.g008

Table 2. Results for all datasets when using half of the subjects–*i.e.*, $D^{(n/2)}$. Reproducibility Scores and average number of biomarkers are shown in the form of mean ± standard deviation. (The caption for Table 1 describes the row ordering).

Name	Average #biomarkers	uRS %	$\widehat{RS}$ %	oRS %
GDS968	1593.67 ± 173.35	26.0 ± 5.87	47.5 ± 3.95	48.6 ± 5.21
GSE7390*	1.38 ± 3.01	0 ± 0	0 ± 0	1.84 ± 2.1
GSE2034*	41.05 ± 144.8	0 ± 0	0 ± 0	6.39 ± 4.72
GSE1456*	58.38 ± 138.11	0 ± 0	0 ± 0	4.5 ± 4.41
GSE11121*	181.8 ± 265.5	0 ± 0	0.09 ± 0.2	9.48 ± 7.67
GDS2546	1266.61 ± 354.72	5.01 ± 7.15	30.8 ± 4.82	33.5 ± 6.3
GDS2545	2051.58 ± 494.77	6.38 ± 7.33	34.0 ± 4.09	38.1 ± 5.92
GDS2547	648.55 ± 255.09	1.82 ± 4.24	23.7 ± 5.86	27.3 ± 6.63
KIPAN	694.48 ± 398.88	0.88 ± 2	12.3 ± 4.91	19.3 ± 8.33
BRCA	36.75 ± 171.15	0 ± 0	0 ± 0	2.6 ± 4.2
GDS2771	457.17 ± 863.64	0.09 ± 0.62	0.32 ± 0.64	7.23 ± 8.48
GDS3966	2976.15 ± 811.92	10.6 ± 7.83	31.7 ± 4.71	37.9 ± 6.44
GDS4185	2.92 ± 40.20	0 ± 0	0 ± 0	0.701 ± 3.11
Metabric	1301.9 ± 504.80	1.97 ± 3.86	18.5 ± 3.86	21.6 ± 7.3
GDS4431	43.35 ± 284.56	0 ± 0	0 ± 0	2.55 ± 3.82
GDS4719	0.36 ± 1.88	0 ± 0	0 ± 0	2.8 ± 5.31
GDS2737	0.79 ± 7.33	0 ± 0	0 ± 0	1.73 ± 2.82
GDS5218	4207.29 ± 1589.79	7.96 ± 4.08	24.0 ± 5.09	31.5 ± 6.15
GSE13429	339.29 ± 170.11	1.73 ± 0.44	1.66 ± 0.44	14 ± 5.04
GSE25103	309.61 ± 78.24	0.13 ± 0.11	0.27 ± 0.14	6.63 ± 6.79
GSE15097	22788.62 ± 23342.37	0.77 ± 0.62	4.9 ± 2.99	16.8 ± 6.85
GSE15096	22393.87 ± 20624.09	0.74 ± 0.60	5.37 ± 2.9	16.1 ± 6.91
GSE25104	309.93 ± 78.14	0.13 ± 0.11	0.27 ± 0.14	6.64 ± 6.8
GSE15826	1.21 ± 5.41	0 ± 0	2.03 ± 0.06	1.21 ± 0.79
GSE18333	0.01 ± 0.1	0 ± 0	0 ± 0	0 ± 0

https://doi.org/10.1371/journal.pone.0252697.t002

biomarker.shinyapps.io/BiomarkerReprod/), to estimate the Reproducibility Score of this proposed set. If those scores (especially the lower bound uRS(D, BD)) are sufficiently high, they can use those biomarkers, confident that they (as a set) are reproducible. But if not, the researchers may want to explore other ways to identify reproducible biomarkers. One obvious approach is to use a larger dataset, with more instances, as we know that RS increases with the size of the dataset $|D|$; see Heuristic 7 and the analysis in Section D.2 in S1 Appendix. Alternatively (or in addition), recall the reproducibility depends on both the dataset, and the *Biomarker Discovery Algorithm* BD; perhaps some other $\text{BD}_{test,p-val,\ MCC}$ would work better here? We could consider modifying all 3 of the components:

- While the MCC methods are designed to reduce the false positives, it is not clear whether they improve RS. Indeed, our experiments (Fig C.1 in S1 Appendix) show that RS values reduce with MCC—which points to using $\text{BD}_{test,p,\ None}$.
- The non-reproducibility problem might be due to the statistical test used. As noted earlier, the t-test implicitly assumes a Gaussian distribution of the features (or the normality of residuals); perhaps another test would work better for some dataset / distribution of instances.
- Finally, they may want to modify the p-value, p, as other values of p may lead to better reproducibility.

4.2 Limitations

This paper provides an effective way to estimate the reproducibility of the biomarkers found from a dataset, using a biomarker discovery algorithm. While the message of this paper is very general, the specific analyses here all used the standard discovery algorithm $\text{BD}_{t,\ 0.05,\ BH}$, to illustrate that the issues apply to the approaches commonly used. Section C in S1 Appendix explores some other discovery methods, that are also based on t-tests. While this use of t-tests implicitly assumes normality of the feature values, nothing in our general approach relies on this specific test—we could use other tests that make other parametric assumptions. (Note that our analysis appears to work effectively even when dealing with some Bernoulli (non-normal) features.) We anticipate the same approach would hold for other tests, such as Mann-Whitney, Wilcoxon or even multivariate analysis–but this is future work. All of our empirical studies dealt with standard datasets, whose outcomes are binary and whose values were either all real values or all categorical values; none had some of each. Our analytic model considers the overlap of biomarkers found from two datasets, of the same size—*e.g.*, we do not consider how the biomarkers obtained from a size-100 dataset, overlap with those from a size-300 dataset. Finally, the recommendations of the previous subsection suggest another future direction: produce the "$\text{BD}'(D, s)$ algorithm" that, given a dataset D and a minimum score $s > 0$, returns the parameters [$test$, p-val, MCC] for a biomarker discovery algorithm $\text{BD}_{test,p-val,\ MCC}$ that would produce a biomarker set whose Reproducibility Score would be at least s–*i.e.*, we expect that $\text{RS}(D, \text{BD}_{test,p-val,\ MCC}(\cdot)) \geq s$, suggesting this level of reproducibility for the proposed biomarkers $\text{BD}_{test,p-val,\ MCC}(D)$.

4.3 Contributions

This paper has several contributions: (1) It motivates, then provides, a formal definition of reproducibility, that can help researchers evaluate a set of purported biomarkers; (2) It provides a pair of algorithms that can accurately bound this "reproducibility score" for a given dataset and biomarker discovery algorithm; (3) It provides empirical evaluation of these algorithms, over 25 different real-world datasets, to demonstrate that they work effectively; and (4) It introduces a freely-available website https://biomarker.shinyapps.io/BiomarkerReprod/ that runs these algorithms on the dataset entered by a user, and biomarker discovery system, which will allow users to easily evaluate the quality of the biomarker set produced. Given how easy it is to use this tool, we hope that future researchers will automatically use it to quickly evaluate the quality of the purported biomarkers, then include these estimates when they publish their biomarkers. We anticipate this analysis may also lead to new biomarker discovery algorithms, to optimize this reproducibility score.

Supporting information

S1 Appendix.
(PDF)

Acknowledgments

The authors all gratefully acknowledge the funding support from Canada's NSERC (Natural Sciences and Engineer Research Council) and the Alberta Machine Intelligence Institute (Amii). We also benefited from the preliminary research work by Patrick Schwaferts and discussions with Sambasivarao Damaraju.

Author Contributions

Conceptualization: Russell Greiner.

Data curation: Amir Forouzandeh.

Formal analysis: Amir Forouzandeh, Alex Rutar, Russell Greiner.

Funding acquisition: Russell Greiner.

Investigation: Amir Forouzandeh, Alex Rutar, Russell Greiner.

Methodology: Amir Forouzandeh, Alex Rutar, Russell Greiner.

Resources: Amir Forouzandeh, Russell Greiner.

Supervision: Russell Greiner.

Validation: Amir Forouzandeh, Alex Rutar, Russell Greiner.

Visualization: Amir Forouzandeh, Alex Rutar.

Writing – original draft: Amir Forouzandeh, Russell Greiner.

Writing – review & editing: Amir Forouzandeh, Alex Rutar, Russell Greiner.

References

1. Strimbu K, Tavel JA. What are biomarkers?Current Opinion in HIV and AIDS. 2010; 5(6):463. https://doi.org/10.1097/COH.0b013e32833ed177 PMID: 20978388
2. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530. https://doi.org/10.1038/415530a
3. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005; 365(9460):671–679. https://doi.org/10.1016/S0140-6736(05)17947-1 PMID: 15721472
4. Blau N. Genetics of phenylketonuria: then and now. Human mutation. 2016; 37(6):508–515. https://doi.org/10.1002/humu.22980 PMID: 26919687

5. Koenig RJ, Peterson CM, Jones RL, Saudek C, Lehrman M, Cerami A. Correlation of glucose regulation and hemoglobin AIc in diabetes mellitus. New England Journal of Medicine. 1976; 295(8):417–420. https://doi.org/10.1056/NEJM197608192950804 PMID: 934240

6. Bush WS, Moore JH. Genome-wide association studies. PLoS computational biology. 2012; 8(12): e1002822. https://doi.org/10.1371/journal.pcbi.1002822 PMID: 23300413

7. Boulesteix AL, Slawski M. Stability and aggregation of ranked gene lists. Briefings in bioinformatics. 2009; 10(5):556–568. https://doi.org/10.1093/bib/bbp034 PMID: 19679825

8. Zou J, Hao C, Hong G, Zheng J, He L, Guo Z. Revealing weak differential gene expressions and their reproducible functions associated with breast cancer metastasis. Computational biology and chemistry. 2012; 39:1–5. https://doi.org/10.1016/j.compbiolchem.2012.04.002 PMID: 22634492

9. Dhami R, Passini MA, Schuchman EH. Identification of novel biomarkers for Niemann–Pick disease using gene expression analysis of acid sphingomyelinase knockout mice. Molecular Therapy. 2006; 13 (3):556–564. https://doi.org/10.1016/j.ymthe.2005.08.020 PMID: 16214420

10. Shlomi T, Cabili MN, Ruppin E. Predicting metabolic biomarkers of human inborn errors of metabolism. Molecular systems biology. 2009; 5(1):263. https://doi.org/10.1038/msb.2009.22 PMID: 19401675

11. Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A. Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. BMC bioinformatics. 2007; 8(1):415. https://doi.org/10.1186/1471-2105-8-415 PMID: 17963508

12. Zacharias HU, Rehberg T, Mehrl S, Richtmann D, Wettig T, Oefner PJ, et al. Scale-invariant biomarker discovery in urine and plasma metabolite fingerprints. Journal of proteome research. 2017; 16 (10):3596–3605. https://doi.org/10.1021/acs.jproteome.7b00325 PMID: 28825821

13. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, et al. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. NeuroImage. 2017; 147:736–745. https://doi.org/10.1016/j.neuroimage.2016.10.045 PMID: 27865923

14. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clinical cancer research. 2007; 13(11):3207–3214. https://doi.org/10.1158/1078-0432.CCR-06-2765 PMID: 17545524

15. Pearl J. Causality. Cambridge university press; 2009.

16. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set?Bioinformatics. 2004; 21(2):171–178. https://doi.org/10.1093/bioinformatics/bth469 PMID: 15308542

17. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001; 98(19):10869–10874. https://doi.org/10.1073/pnas.191367098 PMID: 11553815

18. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nature genetics. 2002; 33(1):49. https://doi.org/10.1038/ng1060 PMID: 12469122

19. Alizadeh AA, Gentles AJ, Alencar AJ, Liu CL, Kohrt HE, Houot R, et al. Prediction of survival in diffuse large B-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. Blood. 2011; 118(5):1350–1358. https://doi.org/10.1182/blood-2011-03-345272 PMID: 21670469

20. Miklos GLG, Maleszka R. Microarray reality checks in the context of a complex disease. Nature biotechnology. 2004; 22(5):615. https://doi.org/10.1038/nbt965 PMID: 15122300

21. Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, et al. Recommendations for biomarker identification and qualification in clinical proteomics. Science translational medicine. 2010; 2 (46):46ps42–46ps42. https://doi.org/10.1126/scitranslmed.3001249 PMID: 20739680

22. Rossing K, Mischak H, Dakna M, Zürbig P, Novak J, Julian BA, et al. Urinary proteomics in diabetes and CKD. Journal of the American Society of Nephrology. 2008; 19(7):1283–1290. https://doi.org/10.1681/ASN.2007091025 PMID: 18448586

23. Haubitz M, Good DM, Woywodt A, Haller H, Rupprecht H, Theodorescu D, et al. Identification and validation of urinary biomarkers for differential diagnosis and evaluation of therapeutic intervention in anti-neutrophil cytoplasmic antibody-associated vasculitis. Molecular & Cellular Proteomics. 2009; 8 (10):2296–2307. https://doi.org/10.1074/mcp.M800529-MCP200 PMID: 19564150

24. Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012; 483(7391):531–533. https://doi.org/10.1038/483531a PMID: 22460880

25. Wen H, Wang HY, He X, Wu CI. On the low reproducibility of cancer studies. National science review. 2018; 5(5):619–624. https://doi.org/10.1093/nsr/nwy021 PMID: 31258951

26. Baker M. Reproducibility crisis? Nature. 2016; 533(26):353–66.

27. Holte RC, Acker L, Porter BW, et al. Concept Learning and the Problem of Small Disjuncts. In: IJCAI. vol. 89; 1989. p. 813–818.

28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

29. Ciucci S, Ge Y, Durán C, Palladini A, Jiménez-Jiménez V, Martínez-Sánchez LM, et al. Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies. Scientific reports. 2017; 7:43946. https://doi.org/10.1038/srep43946 PMID: 28287094

30. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016.

31. Frantz S. An array of problems. Nature Reviews Drug Discovery. 2005; 4:362–363. https://doi.org/10.1038/nrd1746 PMID: 15902768

32. Li M, Hong G, Cheng J, Li J, Cai H, Li X, et al. Identifying reproducible molecular biomarkers for gastric cancer metastasis with the aid of recurrence information. Scientific reports. 2016; 6:24869. https://doi.org/10.1038/srep24869 PMID: 27109211

33. Zhang M, Yao C, Guo Z, Zou J, Zhang L, Xiao H, et al. Apparently low reproducibility of true differential expression discoveries in microarray studies. Bioinformatics. 2008; 24(18):2057–2063. https://doi.org/10.1093/bioinformatics/btn365 PMID: 18632747

34. Ioannidis JP. Biomarker failures. Clinical chemistry. 2013; 59(1):202–204. https://doi.org/10.1373/clinchem.2012.185801 PMID: 22997282

35. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Academy of Sciences. 2006; 103(15):5923–5928. https://doi.org/10.1073/pnas.0601231103 PMID: 16585533

36. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. Nature biotechnology. 2006; 24(9):1151. https://doi.org/10.1038/nbt1239 PMID: 16964229

37. Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nature biotechnology. 2006; 24(9):1162. https://doi.org/10.1038/nbt1238 PMID: 17061323

38. Klebanov L, Qiu X, Welle S, Yakovlev A. Statistical methods and microarray data. Nature biotechnology. 2007; 25(1):25. https://doi.org/10.1038/nbt0107-25 PMID: 17211383

39. Gagno S, D'Andrea MR, Mansutti M, Zanusso C, Puglisi F, Dreussi E, et al. A New Genetic Risk Score to Predict the Outcome of Locally Advanced or Metastatic Breast Cancer Patients Treated With First-Line Exemestane: Results From a Prospective Study. Clinical breast cancer. 2019; 19(2):137–145. https://doi.org/10.1016/j.clbc.2018.11.009 PMID: 30584056

40. Chitpin JG, Awdeh A, Perkins TJ. RECAP reveals the true statistical significance of ChIP-seq peak calls. Bioinformatics. 2019; 35(19):3592–3598. https://doi.org/10.1093/bioinformatics/btz150 PMID: 30824903

41. Pavelka N, Pelizzola M, Vizzardelli C, Capozzoli M, Splendiani A, Granucci F, et al. A power law global error model for the identification of differentially expressed genes in microarray data. BMC bioinformatics. 2004; 5(1):203. https://doi.org/10.1186/1471-2105-5-203 PMID: 15606915

42. Alshawaqfeh M, Bashaireh A, Serpedin E, Suchodolski J. Consistent metagenomic biomarker detection via robust PCA. Biology direct. 2017; 12(1):4. https://doi.org/10.1186/s13062-017-0175-4 PMID: 28143486

43. Zhao SD, Li Y. Score test variable screening. Biometrics. 2014; 70(4):862–871. https://doi.org/10.1111/biom.12209 PMID: 25124197

44. Ma S, Zhang Y, Huang J, Han X, Holford T, Lan Q, et al. Identification of non-Hodgkin's lymphoma prognosis signatures using the CTGDR method. Bioinformatics. 2010; 26(1):15–21. https://doi.org/10.1093/bioinformatics/btp604 PMID: 19850755

45. Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C. Stability indicators in network reconstruction. PloS one. 2014; 9(2). https://doi.org/10.1371/journal.pone.0089815 PMID: 24587057

46. Hua J, Bittner ML, Dougherty ER. Evaluating gene set enrichment analysis via a hybrid data model. Cancer informatics. 2014; 13:CIN–S13305.

47. Shi X, Yi H, Ma S. Measures for the degree of overlap of gene signatures and applications to TCGA. Briefings in bioinformatics. 2015; 16(5):735–744. https://doi.org/10.1093/bib/bbu049 PMID: 25552438

48. Efron B. Bootstrap methods: another look at the jackknife. In: Breakthroughs in statistics. Springer; 1992. p. 569–593.

49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; p. 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

50. Rieger KE, Hong WJ, Tusher VG, Tang J, Tibshirani R, Chu G. Toxicity from radiation therapy associated with abnormal transcriptional responses to DNA damage. Proceedings of the National Academy of Sciences. 2004; 101(17):6635–6640. https://doi.org/10.1073/pnas.0307761101 PMID: 15096622

51. Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. Bioinformatics. 2015; 31(14):2318–2323. https://doi.org/10.1093/bioinformatics/btv157 PMID: 25788628

52. Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast cancer research. 2005; 7(6):R953. https://doi.org/10.1186/bcr1325 PMID: 16280042

53. Schmidt M, Böhm D, Von Törne C, Steiner E, Puhl A, Pilch H, et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer research. 2008; 68(13):5405–5413. https://doi.org/10.1158/0008-5472.CAN-07-5206 PMID: 18593943

54. Chandran UR, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, Liang W, et al. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC cancer. 2007; 7(1):64. https://doi.org/10.1186/1471-2407-7-64 PMID: 17430594

55. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, et al. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. Journal of clinical oncology. 2004; 22(14):2790–2799. https://doi.org/10.1200/JCO.2004.05.158 PMID: 15254046

56. Data generated by the TCGA Research Network: http://cancergenome.nih.gov/;.

57. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nature medicine. 2007; 13(3):361. https://doi.org/10.1038/nm1556 PMID: 17334370

58. Gustafson AM, Soldi R, Anderlind C, Scholand MB, Qian J, Zhang X, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Science translational medicine. 2010; 2(26):26ra25–26ra25. https://doi.org/10.1126/scitranslmed.3000251 PMID: 20375364

59. Xu L, Shen SS, Hoshida Y, Subramanian A, Ross K, Brunet JP, et al. Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases. Molecular Cancer Research. 2008; 6(5):760–769. https://doi.org/10.1158/1541-7786.MCR-07-0344 PMID: 18505921

60. Hutcheson J, Scatizzi JC, Siddiqui AM, Haines III GK, Wu T, Li QZ, et al. Combined deficiency of proapoptotic regulators Bim and Fas results in the early onset of systemic autoimmunity. Immunity. 2008; 28 (2):206–217. https://doi.org/10.1016/j.immuni.2007.12.015 PMID: 18275831

61. Becker AM, Dao KH, Han BK, Kornu R, Lakhanpal S, Mobley AB, et al. SLE peripheral blood B cell, T cell and myeloid cell transcriptomes display unique profiles and each subset contributes to the interferon signature. PloS one. 2013; 8(6):e67003. https://doi.org/10.1371/journal.pone.0067003 PMID: 23826184

62. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346. https://doi.org/10.1038/nature10983 PMID: 22522925

63. Alter MD, Kharkar R, Ramsey KE, Craig DW, Melmed RD, Grebe TA, et al. Autism and increased paternal age related changes in global levels of gene expression regulation. PloS one. 2011; 6(2):e16715. https://doi.org/10.1371/journal.pone.0016715 PMID: 21379579

64. Fernandez DR, Telarico T, Bonilla E, Li Q, Banerjee S, Middleton FA, et al. Activation of mammalian target of rapamycin controls the loss of TCRζ in lupus T cells through HRES-1/Rab4-regulated lysosomal degradation. The Journal of Immunology. 2009; 182(4):2063–2073. https://doi.org/10.4049/jimmunol.0803600 PMID: 19201859

65. Burney RO, Talbi S, Hamilton AE, Vo KC, Nyegaard M, Nezhat CR, et al. Gene expression analysis of endometrium reveals progesterone resistance and candidate susceptibility genes in women with endometriosis. Endocrinology. 2007; 148(8):3814–3826. https://doi.org/10.1210/en.2006-1692 PMID: 17510236

66. Raue U, Trappe TA, Estrem ST, Qian HR, Helvering LM, Smith RC, et al. Transcriptome signature of resistance exercise adaptations: mixed muscle and fiber type specific profiles in young and old adults. American Journal of Physiology-Heart and Circulatory Physiology. 2012;. https://doi.org/10.1152/japplphysiol.00435.2011 PMID: 22302958

67. Venkatachalam R, Verwiel ET, Kamping EJ, Hoenselaar E, Görgens H, Schackert HK, et al. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. International journal of cancer. 2011; 129(7):1635–1642. https://doi.org/10.1002/ijc.25821 PMID: 21128281

68. Peng CH, Liao CT, Peng SC, Chen YJ, Cheng AJ, Juang JL, et al. A novel molecular signature identified by systems genetics approach predicts prognosis in oral squamous cell carcinoma. PloS one. 2011; 6 (8):e23452. https://doi.org/10.1371/journal.pone.0023452 PMID: 21853135

69. Närvä E, Autio R, Rahkonen N, Kong L, Harrison N, Kitsberg D, et al. High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. Nature biotechnology. 2010; 28(4):371. https://doi.org/10.1038/nbt.1615 PMID: 20351689

70. Lee CH, Chang JSM, Syu SH, Wong TS, Chan JYW, Tang YC, et al. IL-1β promotes malignant transformation and tumor aggressiveness in oral cancer. Journal of cellular physiology. 2015; 230(4):875–884. https://doi.org/10.1002/jcp.24816 PMID: 25204733

71. r Pamphlett R. Affymetrix 6.0 study of sporadic motor neuron disease patients and controls., geo, V1.; 2010. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15826.

72. Mao X, Yu Y, Boyd LK, Ren G, Lin D, Chaplin T, et al. Distinct genomic alterations in prostate cancers in Chinese and Western populations suggest alternative pathways of prostate carcinogenesis. Cancer research. 2010; 70(13):5207–5212. https://doi.org/10.1158/0008-5472.CAN-09-4074 PMID: 20516122