Journal Title , 2025, pp. 1–9

doi: DOI
Advance Access Publication Date: Day Month Year
Paper

PAPER

Statistical knockoffs improve biomarker discovery from transcriptomic data

Julie Cartier,[1,2,3,*] Johanna Lagoas,[4] Adeline Fermanian,[5] Chloé-Agathe Azencott[1,2,3] and Florian Massip[1,2,3,*]

[1]Centre for Computational Biology, MinesParis, PSL University, 60 bd Saint-Michel, 75272, Paris, France, [2]Institut Curie, PSL University, 11 rue Pierre et Marie Curie, 75005, Paris, France, [3]U1331, INSERM, 11 rue Pierre et Marie Curie, 75005, Paris, France, [4]AgroParisTech, Paris-Saclay University, 22 place de l'Agronomie, 91123, Palaiseau, France and [5]LOPF, LOPF Califrais'Machine Learning Lab, Paris, France

*Corresponding author. florian.massip@minesparis.psl.eu*Corresponding author. julie.cartier@mineparis.psl.eu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Advances in sequencing technologies have enabled the generation of large amounts of data, offering new possibilities to identify relationships between biological units (*e.g.* genes) and phenotypic traits (*e.g.* disease outcomes). Yet, identifying these associations using variable selection methods remains challenging due to the high dimension ($p \gg n$) and the correlation structure of the data. To address these challenges, we study the applicability of the knockoff (KO) procedure. Introduced by Barber and Candès in 2015, the KO variable selection procedure has shown promising results on real biological data, such as Genome Wide Association Studies. This method seeks to identify the truly important predictors by overcoming the correlation structure between variables while controlling the false discovery rate. Here, we study the applicability of the KO procedure on transcriptomic data in a classification setting. We conduct an extensive simulation study using real transcriptomic data to evaluate the performance of the KO framework in the context of high-dimensional classification. We find that the KO framework outperforms widely used variable selection models, and that using KO aggregation to mitigate the effect of KO stochasticity improves stability while maintaining the same power. Finally applied to three real transcriptomic datasets, the KO framework made very few discoveries, highlighting its conservative nature and suggesting that other methods may substantially overestimate the number of relevant features.

Key words: Knockoffs, Variable selection, Transcriptomic

Introduction

Thanks to the progress of sequencing technologies, it is now possible to collect and access sequencing and genomic information for large cohorts of patients, paving the way for the use of these technologies for many biotechnological and medical applications. In particular, patient stratification and disease grading based on the results of a sequencing experiment is seen as a promising avenue for several diseases, and in cancer in particular [1, 2]. But despite a lot of research on the topic, discovering biomarkers using transcriptomic data remains challenging. In particular, it has often been noted that transcriptomic signatures for a disease are not always reliable [3], difficult to reproduce, and that the genes selected as biomarkers are usually highly unstable [4, 5].

Several reasons make biomarker discovery particularly difficult in this context. First, despite the increase in cohort sizes, the number of explanatory variables remains much larger than the number of patients, which is well known to be a difficult statistical setup (often called "high dimension" or "large p small n") [6, 7]. An additional difficulty arises due to the high level of correlation between the different gene's expression. Indeed, genes that interact to take part to common biological functions are often co-regulated by similar mechanisms, inducing correlation between genes' expressions. As a result, it becomes difficult to differentiate genes that are truly associated to the outcome of interest from those that appear associated to the outcome solely because they are correlated to a causal gene.

Recently, a novel statistical method named Knockoff (KO) [8] has been proposed to tackle the problem of variable selection for correlated features in high dimension. KO have been applied to a range of biological questions, in particular in Genome-Wide Association Studies [9, 10], where it was demonstrated that KO allowed to control the number of false discoveries. However, up to now, only few studies have suggested to apply KO to transcriptomic data [11, 12, 13] and to the best of our knowledge, applying KO to transcriptomic data, in high dimension, for classification, remains unexplored. Hence, whether KO can improve variable selection for transcriptomic data remains to be studied. In addition, because KO is a flexible method, identifying the optimal way to apply it to real data is complex,

© The Author 2025. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

which can prevent its use in practice. For instance, many methods have been developed to generate KO matrices whose power might depend on the specificities of users' data.

Moreover, one particular difficulty with biomarker discovery is that one often obtains unstable results, *i.e.* features selected by the model tend to change a lot with small perturbations of the input data, making it especially hard to transfer results from one cohort of patients to the next. While using KO has been shown to improve false discovery control, the impact of this statistical method on the stability of the selected features is rarely studied. Indeed, as KO generation is non-deterministic, results may vary between several runs of the same method. One way to alleviate this issue is to run the KO framework multiple times and aggregate the results [14, 15, 16, 12].

The goal of this paper is to study whether KO improves variable selection when applied to transcriptomic data for classification tasks. First, we perform a systematic comparison of the performance of different KO generation methods and statistics. We then conduct simulations to compare the KO framework with baselines, namely Lasso regression and Wilcoxon rank-sum tests. We also include KO aggregation in this experiment to assess its performance. Next, we evaluate the stability of the KO framework for variable selection. We focus on different selection scenarios and demonstrate that feature correlation reduces the power of the method. Finally, we apply our resulting KO method on three real-world datasets: two include lung cancer patients and one includes breast cancer patients.

Overview of the KO framework

In this paper, KO refers to the versatile Model-X Knockoffs [17] which we briefly present here. A more detailed summary of the KO procedure can be found in the appendix (see Supp. Mat. A).

Let $\mathbf{y} \in \{0,1\}^n$, $\mathbf{y} = (y_1, \dots y_n)^T$ be the random variables representing the outcomes and $X = (X_{.,1}, \dots, X_{.,p})$ be the matrix of features (here, genes' expressions) such that for any $j \in \{1, \dots p\}$, $X_{.,j} = (X_{1,j}, \dots, X_{n,j})^T \in \mathbb{R}^n$. In the following $X_{.,j}$ will be noted X_j. Let $\mathcal{H}_0$ denote the set of indices of null features such that for all $j \in \mathcal{H}_0, \mathbf{y} \perp\!\!\!\perp X_j | X_{-j}$ where X_{-j} is the matrix X without the jth column. The goal of variable selection is to identify the set of non-null features given the observations X and $\mathbf{y}$, *i.e.*, the set $\mathcal{H}_0^c = \{1, \dots, p\} \backslash \mathcal{H}_0$. The KO framework does so while ensuring that the false discovery rate (FDR) is controlled. Therefore, running the KO framework results in the selection of a set of variables $\hat{\mathcal{S}} \subset \{1, \dots, p\}$ such that:

$$\text{FDR} := \mathbb{E}\left[\frac{|j \in \hat{\mathcal{S}} \cap \mathcal{H}_0|}{\max\left(|\hat{\mathcal{S}}|, 1\right)}\right] \leqslant q, \tag{1}$$

where q is the target FDR level.

The KO framework relies on the construction of false copies of every feature in the data matrix X, which will be used as control variables. These copies are gathered to form the KO matrix $\widetilde{X} = (\widetilde{X}_1, \dots, \widetilde{X}_p) \in \mathbb{R}^{n \times p}$. To guarantee FDR control (that is, Eq. (1)), the KO matrix needs to fulfill two conditions, namely that the distribution of the fake variables is as similar as possible to the distribution of the real features ("swap property"), and that the KO matrix $\widetilde{X}$ is independent of the outcomes $\mathbf{y}$ conditionally on the real feature matrix X ("conditional independence property").

A statistic W_j is then constructed for every feature j based on the knockoff $\widetilde{X}_j$ and the true feature X_j. A high value of W_j must provide evidence that feature j is non-null. Candès et al. [17] have demonstrated that, under some assumptions, the set of features $\hat{\mathcal{S}} = \{j : W_j \geqslant \tau_+\}$, with

$$\tau_+ := \min\left\{t > 0 : \frac{1 + |\{j : W_j \leqslant -t\}|}{|\{j : W_j \geqslant t\}|} \leqslant q\right\}$$

allows to control the FDR.

Additionally to control the FDR, a good variable selection method must ensure good power, that is, must be able to retrieve the non-null features. Mathematically, the power is defined by

$$\text{Power} := \frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0^c|}{|\mathcal{H}_0^c|}.$$

Unlike FDR control, the KO methodology does not guarantee good power by construction and power analysis is theoretically very challenging [18]. Intuitively, to have good power, $\widetilde{X}$ should be as different from the original matrix X as possible (the extreme choice $\widetilde{X} = X$ being a valid choice in terms of FDR control but having a very bad power).

In summary, the task is to build KO matrices and statistics that satisfy the swap and conditional independence properties in order to control the number of false discoveries while achieving the best possible power. Various methods have been proposed to address this problem. Here, we present a few methods that can be used in the context of high-dimensional data and classification.

Generation of KO features

The most common methods assume that the feature matrix X follows a Gaussian distribution. Under this assumption, a parametrized family of distributions of $\widetilde{X}|X$ can be computed, from which a valid KO for the original features can be sampled [17]. Several families of approaches have been proposed under this assumption.

In the family of Mean Absolute Correlation (MAC) Gaussian methods, the KO features are computed to minimize the correlation between a feature X_j and its knockoff $\widetilde{X}_j$ [17]. Among the different MAC approaches, the semidefinite program (SDP) is expected to be the most powerful as p becomes large. The MAC-based KO frameworks suffer from a tendency to identify some KO of non-null original features as important features in the model instead of the non-null features themselves [19]. To address this issue, a new family of Gaussian KO generation methods called Minimum Reconstructability (MRC) has been developed. One typical approach is to maximize the conditional variance $\mathbb{V}(X_j | X_{-j}, \widetilde{X})$, for any $j \in \{1, ..., p\}$. This approach is known as the minimum variance-based reconstructability (MVR) method. Additionnally, conditional independent (CI) KO [18, 20] satisfies conditional independence between a feature X_j and its knockoff $\widetilde{X}_j$, given the other covariates X_{-j}.

Finally, the "Sequential Conditional Independant Pairs" (SCIP) algorithms [17] allow for the construction of valid KO with no further assumption on the distribution of the features X. In practice, a linear empirical version of this algorithm is often used, called the Linear Sequential Conditional Independant Pairs algorithm (LSCIP) [12, 21].

We stress that this list of methods for computing valid knockoffs is not exhaustive. We have focused on methods particularly suitable for transcriptomic-like data. For instance, methods using deep learning models have been proposed [22, 23],

but we will not consider these since our data sets have a small sample size.

KO test statistics

Most KO statistics are built as differences between variable importance metrics. More precisely, they are obtained by applying a machine learning model to $(X, \widetilde{X})$ with target $\mathbf{y}$, choosing a variable importance metric associated to the model, and computing the difference between the importance measure assigned to a feature and that assigned to its KO.

For linear settings, an intuitive option is to use the absolute values of the coefficients of a penalized logistic regression (LASSO Coefficient Difference (LCD) or Elastic-Net penalized logistic regression Coefficient Difference (EN-CD)) [8, 17] as the variable importance. Likewise, variable importance in non-linear models can be used to define KO test statistics. These include differences in Gini impurity in random forests [19], risk reduction in boosting (RRB) in boosted trees [13], and several other boosted tree-based statistics [11], as well as approaches based on deep learning [24].

Finally, some statistics proposed in the literature cannot be directly expressed as a variable difference. One example is the masked likelihood ratio (MLR) [25], which is based on the estimation of log-likelihood ratios. This method accounts for the high correlation between a feature and its KO by lowering the absolute value of their statistic when the correlation is large. Another example is the Lasso Signed Max (LSM) [8].

KO aggregation

The stochastic nature of KO generation methods means that applying the same KO construction method twice to the same data matrix will generate different KO matrices. This results in different subsets of selected variables at each iteration. KO aggregation addresses this issue by applying the KO procedure multiple times and "aggregating knowledge" from these runs to select variables. Several aggregation methods have been proposed, the most recent of which is the Knockoff-π (KOPI) [12] framework.

Materials and Methods

To evaluate the performance of the KO framework on high-dimensional transcriptomic data, we use simulated outcomes computed from real transcriptomic data. This allows us to keep the correlation structure of biological data (and its specificities such as non-Gaussianity) while knowing the true variables to recover.

Real data sets

The CRUKPAP data [26] is composed of healthy volunteers and patients with benign lung condition ($n = 160$) and subjects diagnosed with lung cancer ($n = 199$). For all 369 samples, we have transcriptomic data from nasal epithelium samples (RNA-sequencing, $18,072$ genes) from which 749 genes have been identified as involved in patients' response to smoke. In the following, we keep only those 749 risk genes to form the data matrix (X_{CRUKPAP})

The AEGIS (Airway Epithelial Gene Expression in the Diagnosis of Lung Cancer) data [27] comes from a clinical trial conducted in the USA, Canada and Europe. This data set is composed of 505 clinical patients including 308 patients diagnosed with lung cancer for whom we have gene expression levels (microarrays) of nasal epithelium ($16,127$ genes). We use the set of 535 genes that were found differentially expressed in a subset of the entire cohort. For the experiments with real outcomes, we use cancer status as the outcome for both the CRUKPAP and AEGIS data sets.

Finally, the breast cancer (BC) cohort is composed of gene expression data (RNA sequencing, $18,802$ genes) from tumors of patients in the SCAN-B project (Sweden data canceronome analysis network breast initiative) [28]. These data were used to predict biomarkers in breast cancer, identifying 869 genes with non-zero weights in at least one of five biomarker classifiers. We focus on the estrogen receptor (ER), progesterone receptor (PgR) and human epidermal growth factor receptor 2 (HER2) status. Thus, our data matrix includes expression levels of these 869 genes for 2,827 samples with no missing data concerning the ER, PgR, or HER2 status. ER and PgR status described protein content, with, respectively $2,600$ and $2,458$ positives, while HER2 status reflects gene amplification, with 338 amplified cases.

We use the same pre-processing steps (quality check, batch effect correction, normalization) for these three transcriptomic data sets as in previous studies [26, 27, 28].

Outcome simulation

Unless otherwise specified, the features are standardized to have zero-mean and unit-variance, that is, for $j \in \{1, \dots, p\}$, we apply the operation $X_j \leftarrow (X_j - \overline{X}_j)/\sigma_j$, where $\overline{X}_j$ is the empirical mean of feature j and σ_j its standard deviation.

We denote by k the number of non-zero features, that is, $k = |\mathcal{H}_0^c|$ and set $k = 10$ in all experiments. We construct a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ by randomly sampling k indices from $\{1, \dots, p\}$ to which we assign values drawn uniformly from $\{-1, 1\}$, and set all remaining components of $\boldsymbol{\beta}$ to zero. We then multiply $\boldsymbol{\beta}$ with a magnitude parameter a.

We consider two simulation settings. In the linear simulation setting, for each $i \in \{1, \dots, n\}$, we generate the ith component of the outcome vector $\mathbf{y} \in \{0, 1\}^n$ according to the following generalized linear model:

$$y_i = \begin{cases} 1 \text{ if } \left(1 + e^{-X_{i,\cdot}\boldsymbol{\beta}}\right)^{-1} > 0.5, \\ 0 \text{ otherwise,} \end{cases} \tag{2}$$

where $X_{i,\cdot}$ is the ith row of X. We take $a = 10$, calibrated such that the two classes are well separated for all data sets (see Supp. Mat. B.1: Fig. S1).

In the interaction simulation setting, we assume that p and k are even and define a new $\boldsymbol{\beta} \in \mathbb{R}^{\frac{p}{2}}$ that has $\frac{k}{2}$ non zero components. Without loss of generality, we assume that the columns of the data matrix X, and correspondingly the entries of $\boldsymbol{\beta}$, can be reordered such that the indices of the non-zero features are the first ones. Then, the ith component of the outcome vector $\mathbf{y}$ is defined as follows:

$$y_i = \begin{cases} 1 \text{ if } \left(1 + e^{-f(X_{i,\cdot},\boldsymbol{\beta})}\right)^{-1} > 0.5, \\ 0 \text{ otherwise,} \end{cases} \tag{3}$$

where $f(X_{i,\cdot}, \boldsymbol{\beta}) = \sum_{j=1}^{\frac{k}{2}} \boldsymbol{\beta}_j X_{i,j} X_{i,2j}$ where β_j is the jth coordinate of $\boldsymbol{\beta}$. This allows to model pairwise/quadratic interactions between variables. Note that obtaining well-separated classes in the interaction simulation setting is difficult. We set a to 100 and selected outcome vectors that yielded classes of similar proportions by trial and error (see Supp. Mat. B.1: Fig. S2).

Implementation details

We generate KO matrices and statistics using either the Python package *knockpy* [19] (v1.3.0) or the R package *knockoff* [17] (v0.3.6) except for the LSCIP method and the EN-CD statistic, for which we use the R package *glmnet* [29] (v4.1-7). The parameter $\lambda_{\min}$ refers to the penalization obtained by minimizing the mean cross-validated error which is by default computed with the deviance (difference of log-likelihoods) error over a 10-fold cross-validation. The parameter λ_{oracle} is chosen *a posteriori* and corresponds to the penalization term that maximizes the ratio between true discoveries and false discoveries in a grid of values ranging from 0 to 1 in increments of 0.001. Variable selection with KOPI is performed using the code provided by [12]. The Wilcoxon rank-sum test is applied on already preprocessed data as in [30] with the R *stats* package (v.4.2.0). We correct for multiple hypotheses testing with the Benjamini-Hochberg (BH) [31] procedure. Details are given in supplementary materials (see B.2).

Experimental setups

We denote by M the number of outcome vectors simulated for each experiment. The target FDR level (q) sequence, used by KO-based methods, varies from 0 to 1 in increments of 0.01. For all experiments, unless otherwise specified, we use $X_{\text{CRUKPAP}} \in \mathbb{R}^{369\times 749}$ as data matrix to compute $\mathbf{y}$.

KO generation methods and statistics benchmark

M is set to 100 in the linear simulation setting and to 10 in the interaction simulation setting. From one simulation to the next, only the identity of the genes that are selected to compute $\mathbf{y}$ changes. We compare methods and statistics that seem particularly appropriate for transcriptomic-like data. We include in the comparison the SDP method as the gold standard for KO generation methods, the MVR method, as representative of MRC methods, CI-based KO construction, and the LSCIP algorithm (see Supp. Mat. Algo. 1). We use the LCD statistic for the KO generation methods benchmark since it is the gold standard statistic in the linear simulation setting.

Concerning statistics, we benchmark the LCD, since it outperforms other LASSO-based statistics (see Supp. Mat. C.1.1: Fig. S4). We also consider for linear statistics: the Elastic-net coefficient difference (EN-CD), and the MLR statistic which has shown better performance than the LCD statistic in low-dimensional linear logistic regression [25]. For tree-based methods, the results (see Supp. Fig. S5) show that none of the statistics are particularly effective, but RRB and Gini variable importance seem to be slightly more powerful. However, as RRB is extremely time-consuming to compute, we chose to use Gini-based variable importance. Given the poor performance of deep learning based statistics (see Supp. Fig. S6), we did not include them in the comparison.

Comparison between KO, LASSO, Wilcoxon rank-sum test and KOPI

For $M = 100$ outcome vectors, we run the KO as well as the KOPI procedures with LSCIP and LCD methods. We compare it with the Wilcoxon rank-sum test combined with the BH procedure for target FDR levels in $\{10^{-10}, 5\times 10^{-10}, 10^{-8}, 5\times 10^{-8}, \ldots, 0.01, 0.05, 0.5, 1\}$—this is the equivalent to q in the KO framework. We also add in the comparison a LASSO-penalized logistic regression with λ ranging from 0 to 0.4 in steps of 0.004, $\lambda_{\min}$ and λ_{oracle} (see 3.3). Our metrics to assess the performances are the average power and its standard deviation, as well as the average FDP and its standard deviation.

Stability study

To study the stability of the different variable selection methods with respect to perturbation in the input samples, we use ten times 10-fold cross-validation for a given simulated outcome $\mathbf{y}$. At each step of the cross-validation we apply variable selection procedures to the training set, which contains 90% of the data. We extract the proportion of times each feature has been selected — referred to as the selection frequency — across the 10×10 iterations. We run the entire procedure with 10 different outcomes $\mathbf{y}$.

Real data application

We apply KOPI with LSCIP and LCD methods for different target FDR levels ($q \in \{0.1, 0.2, 0.3, 0.5\}$) to the three real data sets 3.1.

Results

In the following, we present the results of the simulation study conducted to investigate the applicability of the KO framework to transcriptomic data.

KO generation methods and statistics benchmark

We first compare the performance of different KO generation methods and statistics in the linear simulation setting.

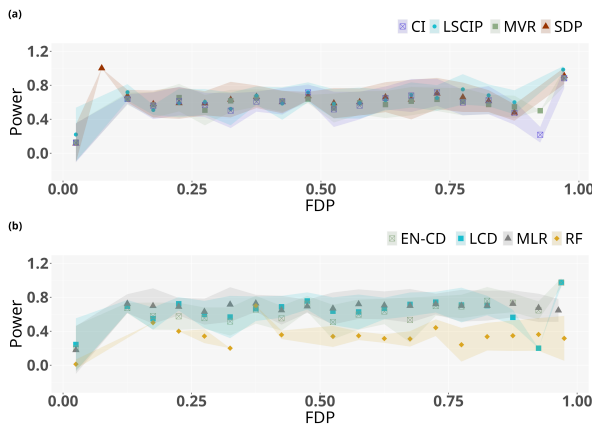

Figure 1. Comparison of the KO generation methods (a) and of different KO statistics (b). Outcomes are simulated from the CRUKPAP feature matrix using the linear simulation setting with $k = 10$ and $M = 100$ repetitions. Power values are given as means over FDP bins of width 0.05. (a) Important variables are selected using the LCD statistics for all KO generation methods. CI: Conditional independence, LSCIP: Linear Sequential Conditional Independant Pairs algorithm, MVR: Minimum Variance-based reconstruction, SDP: semi-definite programming, see section 2.1 for details. (b) KO features are generated using the LSCIP method for all statistics. EN-CD: Elastic Net Coefficients Difference, LCD: Lasso Coefficients Difference, MLR: Maximum Likelihood Ratio, RF: Random Forest, see section 2.2 for details.

We find that MVR, SDP, CI and LSCIP have similar power, as illustrated in Fig. 1. Concerning the statistics, Fig. 1 clearly show better power for linear statistics (LCD, EN-CD, and

MLR)—which was expected as we used a linear model to generate $\mathbf{y}$. Over the 100 replicates, see supplementary figures S3 (c-d)), the EN-CD statistics have a lower power on average than the other two linear statistics for target FDR levels lower than 0.5. Our results also highlight that the EN-CD statistic is conservative with FDP lower on average than the target FDR level and the other three methods for $q < 0.5$. On average, the four KO generation methods and statistics allow control of the number of false discoveries (see Supp. Fig. S3 (a)).

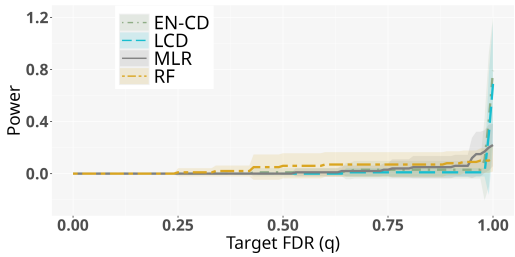

Figure 2. Comparison of different KO statistics in the interaction simulation setting. Outcomes are simulated from the CRUKPAP feature matrix with $k = 10$ and $M = 10$ repetitions. Power values are given as means over FDP bins of width 0.05. EN-CD: Elastic Net Coefficients Difference, LCD: Lasso Coefficients Difference, MLR: Maximum Likelihood Ratio, RF: Random Forest, see section 2.2 for details.

We also compare the performance obtained with the different statistics in the interaction case, to account for the interacting nature of biological processes. All methods, even the RF nonlinear statistic, demonstrate a clear lack of performance (see Fig. 2 and Supp. Mat. C.1.3).

We then conducted the same experiments on the AEGIS and BC datasets and found similar results (Supp. Mat. D.1 and D.2), although we note that the performance of the KO methods are generally better on the BC cohort than on the others. All our analyses, except for the BC data, exhibit large standard deviation intervals (Fig. 1, Supp. Fig. S3 (b-d), and Supp. Fig. S15 and Fig. S16), indicating high variability in our results, which motivates the stability study below.

To conclude, this simulation study shows that all KO generation methods produce similar results, in contrast to other settings in which MVR or CI were shown to outperform SDP [19, 20]. Concerning the choice of statistics, MLR and LCD have shown the best performance in the linear simulation setting. In what follows, we will use the LSCIP method for KO generation because it does not make assumptions about the data structure and the LCD statistic since it is the most widely used in the literature.

After determining which KO generation method and which statistic to use, we compared the performance of the KO framework using the aggregation method KOPI, and found that KOPI is as powerful as the standard KO procedure, and thus include KOPI in the following experiments (see Supp. Fig. S11 and Tab. 1).

Comparison to baselines

We investigate how KO methods compare to other standard variable selection procedures. We consider the linear simulation setting and compare KO methods to a LASSO-penalized logistic regression and Wilcoxon rank-sum tests (which has shown a promising reduction in of the number of false discoveries compared to some other methods for large sample sizes in differential expression analysis [30, 32]).

Method	Parameter	FDP (mean)	Power (mean)
KO	$q = 0.2$	0.13 *(sd: 0.15)*	0.47 *(sd: 0.31)*
	$q = 0.5$	0.48 *(sd: 0.27)*	0.65 *(sd: 0.14)*
KOPI	$q = 0.2$	0.03 *(sd: 0.07)*	0.41 *(sd: 0.29)*
	$q = 0.5$	0.35 *(sd: 0.22)*	0.66 *(sd: 0.17)*
Wilcoxon	$\alpha = 10^{-6}$	0.91 *(sd: 0.21)*	0.60 *(sd: 0.23)*
	$\alpha = 0.05$	0.98 *(sd: 0.01)*	0.83 *(sd: 0.16)*
LASSO	λ_{oracle}	0.37 *(sd: 0.31)*	0.28 *(sd: 0.23)*
	$\lambda_{\min}$	0.88 *(sd: 0.02)*	0.76 *(sd: 0.13)*

Table 1. Comparison of the power and FDP obtained with different variable selection methods: the KO procedure and KOPI with $q = 0.2$ and $q = 0.5$, the LASSO with penalization parameters $\lambda_{\min}$ and λ_{oracle}, and the Wilcoxon rank-sum test with $\alpha = 10^{-6}$ and $\alpha = 0.05$. The mean and the standard deviation (sd) are computed for $M = 100$ linearly simulated $\mathbf{y}$ on the CRUKPAP features matrix.

Table 1 summarizes the power and FDP values obtained for the different methods. It underlines that the Wilcoxon rank-sum test and the LASSO computed with $\lambda_{\min}$ induce many false discoveries with an average FDP above 0.85. The two KO-based methods demonstrate better power over FDP ratio than other methods. In particular, we optimize *a posteriori* the LASSO penalization parameter to maximize the power over FDP ratio (called λ_{oracle}), and found that KO-based methods outperformed the LASSO even in the LASSO most ideal case. As in our previous benchmarks, standard deviation values in Tab. 1 illustrate high variability.

To confirm our results, despite large standard deviation for both power and FDP, we compare the selection performance for larger sequences of false discoveries control parameters for the KO, KOPI and Wilcoxon rank-sum test and a sequence of penalization parameters for the LASSO. Regardless of the FDP interval considered after selection, the KO procedure and KOPI are more powerful than the LASSO and the Wilcoxon rank-sum test and show similar power (see Supp. Fig. S10 and Fig. S11).

Overall, these comparisons with baselines models demonstrate the ability of KO-based methods to improve selection power and to control the number of false discoveries.

Stability study

Our previous results show that the KO framework controls the number of false discoveries while maintaining good power in the linear simulation setting. However, these performance metrics do not give any information about the impact of the stochasticity of the KO framework and the confidence we can have in the resulting selected subsets of genes. To fill this gap, we now turn to the study of feature selection stability that we use as a measure of reliability. Stability is defined as the extent to which two subsets selected by the same method resemble each other when the sample to which it is applied is disturbed [4]. To study variable selection stability, we simulate perturbation by subsampling of the initial samples (see section 3.4.3 for details) and analyze the selection frequency of each variable.

The results are given in Fig. 3. We find that LASSO-penalized logistic regressions, with penalization parameter

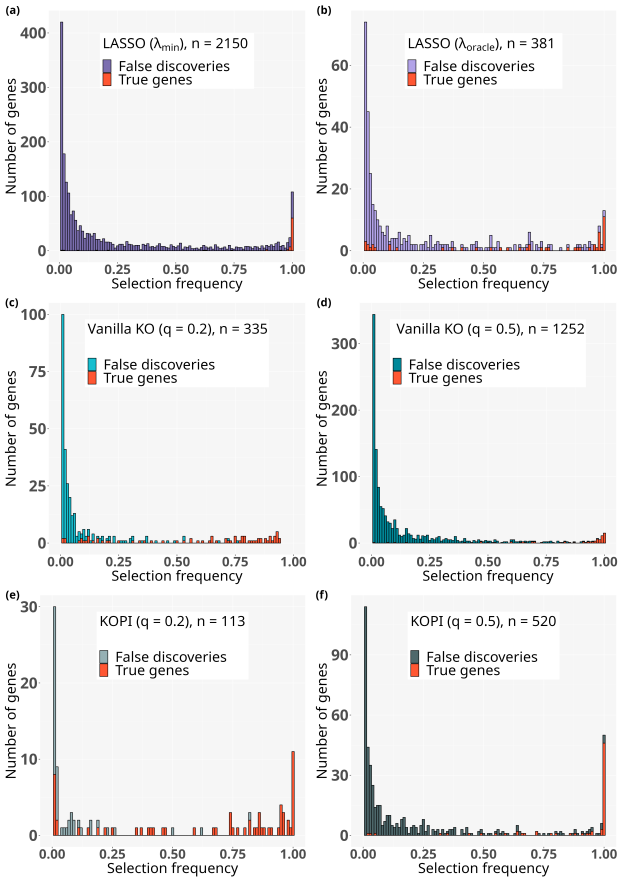

Figure 3. Histograms of the frequency of features selected *more than once* in different variable selection frameworks. Selection frequencies are computed over 10 iterations of ten-fold subsampling for each method and simulated outcome, in the linear simulation setting with $k = 10$ and the CRUKPAP features matrix. n gives the number of features among the 749×10 selected at least once across the 100×10 iterations. Features that were never selected across any iterations are not displayed. (a-b): Selection performed with LASSO-penalized logistic regressions, (a) for $\lambda = \lambda_{min}$ or (b) $\lambda = \lambda_{oracle}$ (c-d): Selection performed with knockoffs for FDR level $q = 0.2$ (c) and q = 0.5 (d). (e-f): Selection performed with KOPI with target FDR level $q = 0.2$ (e) and $q = 0.5$ (f).

chosen by cross-validation, selects too many variables and is unstable. In particular, out of a total of 749×10 genes, it has $2,070$ alternatively selected genes, *i.e.*, genes that are neither always nor never selected–in other words, their selection frequency is not equal to 0 or 1. In addition, among the set of systematically selected genes, more than 40% are false discoveries.

Conversely, knockoff-based methods tend to limit the selection of alternatively selected features, 335 for the vanilla knockoff and 102 for KOPI with $q = 0.2$. Taking $q = 0.5$ and thus relaxing the constraint on the number of false discoveries enables the selection of true causal features at almost each iteration but also leads to an increase of the number of alternatively selected features (1 237 with vanilla knockoff and 470 with KOPI). KOPI leads to fewer alternatively selected genes as the KO procedure, showing the ability of the aggregation scheme to mitigate the effect of knockoff stochasticity while maintaining similar power (see Fig. 3 and Supp. Fig. S11.

Finally, the stability of the KOPI framework is equivalent to that of the oracle LASSO (set *a posteriori* to maximize the true-to-false discovery ratio), showing that KOPI is as good as the best possible selection set obtained with the lasso.

Method	Parameter	TD (FD) (freq = 1)	TD (FD) (freq > 0)
KO	$q = 0.2$	0 (0)	74 (261)
	$q = 0.5$	15 (0)	75 (1177)
KOPI	$q = 0.2$	11 (0)	64 (49)
	$q = 0.5$	46 (4)	73 (447)
LASSO	λ_{oracle}	11 (2)	52 (329)
	$\lambda_{\min}$	60 (48)	80 (2070)

Table 2. Number of true discoveries (TD) and false discoveries (FD) among systematically selected features (selection frequency = 1) and features selected at least once (freq > 0) for different variable selection methods. The selection frequency is computed over 10 iterations of ten-fold subsampling for each method and simulated outcome $\mathbf{y}$. The outcomes $\mathbf{y}$ are simulated in the linear simulation setting with $k = 10$ and the CRUKPAP feature matrix.

.

Table 2 summarizes the number of true and false discoveries found in Fig. 3. As expected, we find that the LASSO makes more true discoveries at the cost of a high number of false discoveries (proportion of true discoveries: 3.7%) while KOPI selects fewer features in order to control the FDR level (see Tab. 2). KOPI also demonstrates a higher proportion of true discoveries among the set of selected genes (for $q = 0.2$, 56.6% and for $q = 0.5$, 14.0%) than the LASSO (0.04%) and the oracle LASSO (13.7%).

Overall, our results illustrate both the conservative nature and the ability of knockoff-based methods to remove false discoveries specifically among systematically selected features. We also find that features selected by KOPI are more reliable than those selected by the KO framework or the LASSO.

Influence of the correlation between features on performance

To study the causes of the differences in performance of the KO method between the different datasets (see Supp. Mat. D and Fig. 1), we compare the correlation matrices of the three cohorts (see Supp. Fig. S19) and find that the matrix data X_{BC} has a less correlated structure than both X_{CRUKPAP} and X_{AEGIS}. This suggests that correlation between causal genes may reduce power.

To test this hypothesis, we introduce difficulty categories ("easy", "medium", and "hard") based on power criteria, which depict how easy it is for the KO framework to find the 10 true causal variables. We categorize in Tab. 3 each of the 100 simulations used in Section 3.4.1 (for additional details and Supp. Mat. D). Tab. 3 shows that the selection is easier with the breast cancer cohort (54 "easy" cases) than with the other two cohorts (0 and 7 "easy" cases for CRUKPAP and AEGIS cohorts).

As an illustration, we compare (see Fig. 4) the average power obtained over ten "easy" settings and ten "hard" settings for the CRUKPAP and the breast cancer cohorts. These curves point out that the CRUKPAP results demonstrate larger variability in performance than the breast cancer results. It also shows that the identity of the 10 non null features is the main source of performance variability. To quantify how correlation

Data sets \ Power	$\geqslant 0.7$ "easy"	$\in]0.25, 0.7[$ "medium"	$\leqslant 0.25$ "hard"
CRUKPAP	0	68	32
AEGIS	7	67	26
Breast cancer	54	43	3

Table 3. Classification of the performance of variable selection using the KO framework for different transcriptomic data sets. The outcomes $\mathbf{y}$ are simulated 100 times with, at each iteration, 10 randomly selected non null genes from a real feature matrix X ($X \in \{X_{\text{CRUKPAP}}, X_{\text{AEGIS}}, X_{\text{BC}}\}$ see Supp. Mat. D for details). The power is given as the average power obtained over an FDP interval: $\mathbf{FDP} \in [0, 0.3]$.

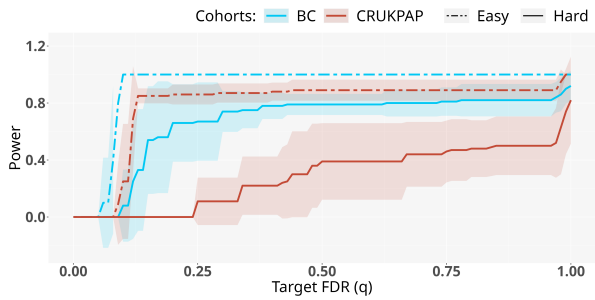

Figure 4. Illustration of the performance variation of the KO procedure (LSCIP + LCD) The power is obtained for the 10 most difficult (continuous line) or the 10 easiest case (dotted line) out of $M = 100$ different outcomes simulations using the linear simulation setting with $k = 10$ and from the CRUKPAP (red) or the breast cancer (blue) feature matrix

affects performance, we show in Fig. 5 the mean power as a function of the average pairwise correlation between the non null features. We find that the highest power values are reached for low pairwise correlations (< 0.2). It also depicts a general tendency of the KO framework to be more powerful on average for lower pairwise correlations. Ultimately, higher correlation in the data matrix makes selection harder in the KO framework. We have shown that selection performance and variability depend partly on the correlation between true causal features.

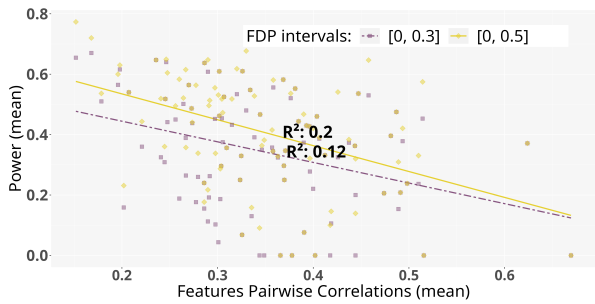

Figure 5. KO variable selection power depending on the pairwise correlation between the 10 real important features. Simulations are conducted with the linear simulation setting on the CRUKPAP feature matrix. The mean power is computed for two different FDP intervals: $[0, 0.3]$ (purple dots) and $[0, 0.5]$ (yellow dots). Each dot corresponds to a different simulation experiments. Continuous and dotted lines represent the linear regression fit of the mean power versus pairwise correlation for the two different FDP intervals.

Application to real data

To evaluate the applicability of KO-based procedures in real settings, we apply KOPI to perform variable selection with real transcriptomic data from CRUKPAP, AEGIS and the breast cancer data sets, but this time $\mathbf{y}$ is set to real outcomes—for which we thus do not know the true causal genes. For CRUKPAP and AEGIS, we aim to select variables involved in cancer diagnosis (cancer/no cancer). In the breast cancer study, we are looking for genes that can discriminate breast cancer biomarkers: ER status (positive/negative), PgR status (positive/negative), and HER2 status (amplified/normal).

We find that for lung cancer predictions, ER and PgR status predictions, no genes are selected with $q = \{0.1, 0.2, 0.3\}$. For q set to 0.5, only 3 genes are selected in the CRUKPAP cohort and 7 in the AEGIS cohort (see Supp. Tab. S1). For the prediction of ER status in the breast cancer cohort, only one gene, the ESR1 gene (coding for the estrogen receptor) is selected. Likewise, in the sets of genes selected (6 genes for $q = 0.3$ and 13 genes for $q = 0.5$) as involved in HER2 status, we retrieve the gene coding for the HER2 protein (ERBB2). The fact that no genes are selected for the PgR status prediction is consistent with the results of the original breast cancer study where the PgR status was shown to be the hardest to predict from PGR expression level [28].

Finally, we also note that the ERBB2 gene is less correlated to other genes as compared to ESR1 and PGR (see Supp. Fig. S20 and Fig. S21). This lower level of correlation might in part explain why the KO-framework selects more genes in the HER2 prediction case, in good agreement with our results of Section 4.4. Overall, these results illustrate the conservative nature of KOPI in real settings.

Discussion

In this paper, we studied the applicability of the KO framework to variable selection problems on transcriptomic data. We focused on a common use case where the goal is to find genes associated with a phenotype of interest—such as a disease. This is in contrast to most applications of KO which are conducted in a regression setting. We benchmark different statistics and KO generation methods, and found that all KO generation methods yield similar results in our simulations. MLR and LCD statistics showed the best performance in the linear simulation setting while being, like other statistics, powerless in the interaction simulation setting. Our extensive simulation study shows that KO succeed in significantly reducing the false discovery rate compared to standard feature selection procedures. While using KO can also reduce statistical power, the KO framework exhibits an improved power over false discovery ratio as compared to standard approaches in our simulations. Finally, we also show that using KO aggregation strategies allows to mitigate the instability of gene selection, a problem often disregarded when comparing feature selection procedures, but nevertheless of major importance.

Applying KO to real transcriptomic data also demonstrates the limitations of the framework. First, the KO framework tends to be very conservative, leading in several of the real examples we studied to very few or sometimes no single discovery. However, these real life examples are taken from previous research where the authors select tens or hundreds of genes to construct clinical classifiers that achieve good performance, suggesting that there is signal in the data and features to be selected. While highlighting the conservative nature of the KO

procedure, this discrepancy also indicates that the genes selected in the original studies potentially contain many false discoveries.

Second, we show that while KO performs well when features are moderately correlated, its power decreases when the correlation between the causal features reaches very high values. While KO are constructed precisely to perform variable selection for correlated features, our analyses show this is effective only up to a point. Interestingly, we find that the level of correlation between the true causal features impacts the performance of the KO selection procedure more strongly than the correlation between causal features and non-causal features.

Third, we focused mostly on the case of $k = 10$ true causal features in our simulations. Qualitatively, we obtain similar results with $k = 30$ (*i.e.* KO still outperforms other feature selection methods, see Supp. Fig. S14). However, we note that the performance of the KO methods deteriorates when the number of causal features increases, see Supp. Fig. S13. Similarly, how the KO framework behaves for larger values of the number of features p and of samples n remains to be investigated. One expected difficulty lies in the computational cost of the generation of the KO matrix, which increases rapidly as the dataset size increases [17].

Finally we find that variable selection performance collapses when the link between gene expression and outcome is not linear. It is however very likely that for many biological processes, the dependency between features and the outcome is not linear. This difficulty is not specific to KO, as all of the classical methods we compare have very low performances in our simulations (see Supp. Fig.S12). Our results thus highlight the need for novel methodological developments to tackle feature selection for correlated data in nonlinear classification settings.

Another interesting result of our study is to underline the instability of classical gene selection procedures, as previously observed [4]. While standard KO do not improve selection stability—notably because of the inherent stochasticity of the KO framework—we find that using KO aggregation methods such as KOPI [12] greatly improves the stability of the selection procedure (Fig. 3), while maintaining similar statistical power. We note however that KOPI calibration is not ideal and the FDR control appears very conservative (see Tab. 2), such that very few discoveries are obtained with a target FDR rate of 0.2. In our simulations, we obtain a low number of false discoveries ($< 5\%$) at the cost of limited power (41%) for a target rate of 0.2. Yet, these results demonstrate a better power-to-FDP ratio than other approaches.

Overall, our study further demonstrates the difficulty of classification tasks with transcriptomic data. We nevertheless show that KO can constitute an interesting strategy, in particular to improve gene selection stability. Our study shows how KO can be applied to transcriptomic dataset, leading the way to more sophisticated approaches. One approach could be to group genes that belong to similar biological pathway, a strategy that was already implemented successfully many times [33], but never in combination with KO.

Keypoints

- Knockoffs-based methods demonstrate a better power-to-false discovery proportion ratio than other commonly used variable selection methods in the context of high-dimensional classification from transcriptomic data.
- The stability of the selection is improved by using an aggregation scheme that mitigates the effects of knockoffs stochasticity while maintaining the same power.
- Both real-world applications and simulations highlight several limitations of the KO framework: its conservative nature, its sensitivity to the correlation structure and the number of true causal features, and its poor performance in nonlinear classification settings.

Competing interests

No competing interest is declared.

Author contributions statement

A.F., C.A.A and F.M. conceived the experiment(s), J.C. and J.L. conducted the experiment(s), and J.C., A.F., C.A.A and F.M. analysed the results. J.C. wrote the manuscript, A.F., C.A.A and F.M. reviewed the manuscript.

Data Availability

The code for all experiments is available in the following Github repository (`https://github.com/Julie-cartier/RNAKnockoffs`). The breast cancer gene expression data are available from the NCBI Gene Expression Omnibus under accession GSE96058 . The AEGIS gene expression data are also available in the NCBI Gene Expression Omnibus under accession number GSE80796.

Acknowledgments

This work was carried out with the financial support of Carnot M.I.N.E.S, as part of its leading project CARINGS project, dedicated to health engineering. This work was also supported by the French Agence Nationale de la Recherche (ANR-19-P3IA-0001). The authors would like to thank the Cancer Research UK, Royal Papworth Hospital and the co-authors of De Biase et al [26] for access to the CRUKPAP cohort data (EGA project number: EGAD50000000333).

References

1. M. Uhlen, C. Zhang, S. Lee, and et al. A pathology atlas of the human cancer transcriptome. *Science*, 357(6352), 2017.
2. O. Govaere, S. Cockell, D. Tiniakos, and et al. Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis. *Science Translational Medicine*, 12(572), 2020.
3. L. Jacob and T. Speed. The healthy ageing gene expression signature for Alzheimer's disease diagnosis: a random sampling perspective. *Genome Biology*, 19(97), 2018.
4. A. C. Haury, P. Gestraud, and J. P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*, 6(12), 2011.
5. L. Ein-Dor, I. Kela, G. Getz, D. Givol, and et al. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2004.
6. IM. Johnstone and DM. Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A*, 367(1906):4237–4253, 2009.

7. R. Clarke, H. Ressom, A. Wang, and et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8:37–49, 2008.
8. R. F. Barber and E. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.
9. M. Sesia, C. Sabatti, and E. Candès. Gene hunting with knockoffs for hidden markov models. *Biometrika*, 106(1):1–18, 2019.
10. M. Sesia, S. Bates, E. Candès, and et al. False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences*, 118(40), 2021.
11. T. Jiang, Y. Li, and A. A. Motsinger-Reif. Knockoff boosted tree for model-free variable selection. *Bioinformatics*, 37(7):976–983, 2021.
12. A. Blain, B. Thirion, O. Grisel, and et al. False discovery proportion control for aggregated knockoffs. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 3417(12), 2023.
13. A. Shen, H. Fu, K. He, and et al. False discovery rate control in cancer biomarker selection using knockoffs. *Cancers*, 11(6):744, 2019.
14. T.-B. Nguyen, J.-A. Chevalier, B. Thirion, and et al. Aggregation of multiple knockoffs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7283–7293, 2020.
15. Z. Ren, Y. Wei, and E. Candès. Derandomizing knockoffs. *Journal of the American Statistical Association*, 118(542):948–958, 2021.
16. Z. Ren and R. F. Barber. Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):122–154, 2023.
17. E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.
18. J. Liu and P. Rigollet. Power analysis of knockoff filters for correlated designs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
19. A. Spector and L. Janson. Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1):252–276, 2022.
20. Z. T. Ke, J. S. Liu, and Y. Ma. Power of knockoff: The impact of ranking algorithm, augmented design, and symmetric statistic. *Journal of Machine Learning Research*, 25(3):1–67, 2024.
21. A. Blain, B. Thirion, J. Linhart, and et al. When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs. *arXiv*, 2407.06892, 2024.
22. J. Jordon, J. Yoon, and M. van der Schaar. KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. *International Conference on Learning Representations*, 2019.
23. Y. Romano, M. Sesia, and E J. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.
24. Y. Lu, Y. Fan, J. Lv, and et al. DeepPINK: reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
25. A. Spector and W. Fithian. Asymptotically optimal knockoff statistics via the masked likelihood ratio. *arXiv*, 2212.08766, 2022.
26. M. S. De Biase, F. Massip, T. T. Wei, and et al. Smoking-associated gene expression alterations in nasal epithelium reveal immune impairment linked to lung cancer risk. *Genome Medicine*, 16(1):54, 2024.
27. J. Perez-Rogers, J. Gerrein, C. Anderlind, and et al. Shared gene expression alterations in nasal and bronchial epithelium for lung cancer detection. *Journal of the National Cancer*, 109(7), 2017.
28. C. Brueffer, J. Vallon-Christersson, D. Grabaut, and et al. Clinical value of rna sequencing–based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter sweden cancerome analysis network—breast initiative. *JCO Precision Oncology*, 2:1–18, 2018.
29. J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
30. Y. Li, X. Ge, F. Peng, and et al. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biology*, 23(79), 2022.
31. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
32. B P. Hejblum, K. Ba, R. Thiébaut, and et al. Neglecting the impact of normalization in semi-synthetic RNA-seq data simulations generates artificial false positives. *Genome Biology*, 25(1):281, October 2024.
33. E O. Paull, A. Aytes, S J Jones, and et al. A modular master regulator landscape controls cancer transcriptional identity. *Cell*, 184(2):334–351.e20, 2021.