Signal Processing 231 (2025) 109894

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

The terminating-random experiments selector: Fast high-dimensional variable selection with false discovery rate control ☆

Jasin Machkour [a],*, Michael Muma [a], Daniel P. Palomar [b]

[a] Robust Data Science Group at Technische Universität Darmstadt, Germany
[b] Convex Optimization Group, Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Keywords:
T-Rex selector
False discovery rate (FDR) control
High-dimensional variable selection
Martingale theory
Genome-wide association studies (GWAS)

ABSTRACT

We propose the Terminating-Random Experiments (T-Rex) selector, a fast variable selection method for high-dimensional data. The T-Rex selector controls a user-defined target false discovery rate (FDR) while maximizing the number of selected variables. This is achieved by fusing the solutions of multiple early terminated random experiments. The experiments are conducted on a combination of the original predictors and multiple sets of randomly generated dummy predictors. A finite sample proof based on martingale theory for the FDR control property is provided. Numerical simulations confirm that the FDR is controlled at the target level while allowing for high power. We prove that the dummies can be sampled from any univariate probability distribution with finite expectation and variance. The computational complexity of the proposed method is linear in the number of variables. The T-Rex selector outperforms state-of-the-art methods for FDR control in numerical experiments and on a simulated genome-wide association study (GWAS), while its sequential computation time is more than two orders of magnitude lower than that of the strongest benchmark methods. The open source R package TRexSelector containing the implementation of the T-Rex selector is available on CRAN.

1. Introduction and motivation

Determining the set of active signals or variables is crucial, e.g., in detection [1–3], antenna array processing [4], distributed learning [5], portfolio optimization [6], and robust estimation [7–11]. In this work, we focus on genome-wide association studies (GWAS) [12], where only a few common genetic variations called single nucleotide polymorphisms (SNPs) among potentially millions of candidates are associated with a phenotype (e.g., disease) of interest [12]. To enable reproducible discoveries, it is essential that (i) the proportion of falsely selected variables among all selected variables is low while (ii) the proportion of correctly selected variables among all true active variables is high. The expected values of these quantities are referred to as the false discovery rate (FDR) and the true positive rate (TPR), respectively. Without FDR control, expensive functional genomics studies and biological laboratory experiments are wasted on researching false positives [13–16].

Establishing FDR control in high-dimensional settings is challenging and, unfortunately, established FDR-controlling methods for low-dimensional data, e.g., [17–19], do not apply to high-dimensional settings. In recent years, the *model-X* knockoff method [20] and derandomized versions thereof [21,22] have been proposed. However, they are computationally demanding. In fact, creating knockoff predictors that mimic the covariance structure of the original predictors renders them infeasible for settings beyond a few thousand variables (see Fig. 1). Moreover, the original derandomized knockoffs approach controls the conservative per family error rate (*PFER*) and the k-family-wise error rate (k-*FWER*) but does not consider the less conservative FDR metric [21]. Only the derandomized approach based on e-values controls the FDR [22]. Nevertheless, the need for running the *model-X* knockoff method multiple times renders both derandomized knockoffs approaches practically infeasible for large-scale high-dimensional settings.

Alternative FDR-controlling approaches that rely on conditional randomization test (*CRT*) p-values [20] are computationally significantly more demanding than the *model-X* knockoff methods (see [20] for a discussion), which renders them infeasible in even relatively small settings (i.e., $p \approx 1000$ candidate variables).

☆ The work of the first author has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 425884435. The work of the second author has been funded by the LOEWE initiative (Hesse Germany) within the emergenCITY center and is supported by the ERC Starting Grant ScReeningData (Project Number: 101042407). The work of the third author has been funded by the Hong Kong GRF 16206123 research grant.

* Corresponding author.
E-mail addresses: jasin.machkour@tu-darmstadt.de (J. Machkour), michael.muma@tu-darmstadt.de (M. Muma), palomar@ust.hk (D.P. Palomar).

https://doi.org/10.1016/j.sigpro.2025.109894
Received 10 August 2024; Received in revised form 20 December 2024; Accepted 8 January 2025
Available online 13 January 2025

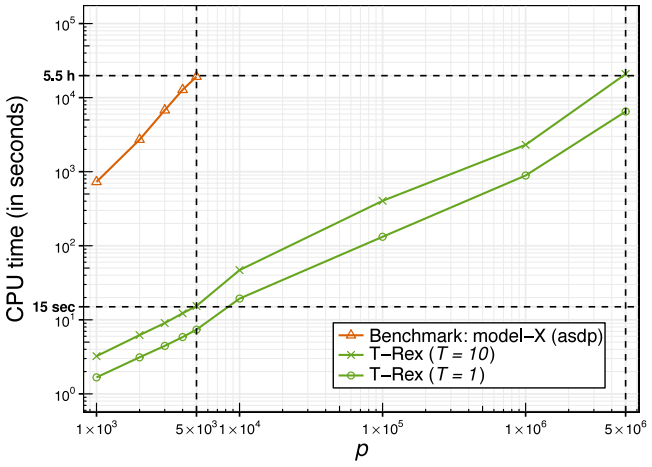


Fig. 1. The sequential computation time of the *T-Rex* selector is multiple orders of magnitude lower than that of the *model-X* knockoff method. Note that, e.g., for $p = 5000$ variables the absolute sequential computation time of the *T-Rex* selector for $T = 10$ included dummies is only 15 s as compared to more than 5.5 h for the *model-X* knockoff method. Moreover, the sequential computation time of the *T-Rex* selector for 5,000,000 variables is comparable to that of the *model-X* knockoff method for only 5000 variables. Note that both axes are scaled logarithmically. Setup: $n = 300$ (observations), $p_1 = 10$ (true active variables), $L = p$ (generated dummies), $K = 20$ (random experiments), SNR $= 1$, $MC = 955$ (Monte Carlo replications) for $p \leq 5000$ and $MC = 100$ for $p > 5000$.

Related lines of research on error-controlled high-dimensional variable selection are centered around stability selection methods [23,24], data-splitting methods [25–28], and post-selection inference [29–32].

In this work, we propose the Terminating-Random Experiments (*T-Rex*) selector, a scalable framework (see Section 2.3) that turns forward variable selection methods into FDR-controlling methods. The *T-Rex* selector fuses the solutions of K early terminated random experiments, in which original and dummy variables compete to be selected in a forward variable selection process. It utilizes dummies in a fundamentally different manner than existing methods (e.g., [33–35]), which are not designed for FDR control. The *T-Rex* calibration algorithm automatically determines its parameters, i.e., (i) the number of generated dummies L, (ii) the number of included dummies before terminating the random experiments T, and (iii) the voting level in the fusion process, such that the FDR is controlled at the target level.

Our main theoretical results are summarized as follows:

1. Using martingale theory [36], we provide a finite sample FDR control proof (Theorem 1) that applies to low- ($p \leq n$) and high-dimensional ($p > n$) settings.
2. We prove that, for the *T-Rex* selector, the dummies can be sampled from any univariate distribution with finite mean and variance (Theorem 2). This is a fundamentally new result, and it does not hold for, e.g., knockoff methods [19,20] that require mimicking the covariance structure of the predictors, which is computationally expensive (see Figure 7 in the supplementary materials.
3. We also prove that the proposed calibration algorithm is optimal in the sense that it maximizes the number of selected variables while controlling the FDR at the target level (Theorem 3).

The major advantages compared to existing methods are:

1. The computation time of the *T-Rex* selector is multiple orders of magnitude lower compared to that of the current benchmark method (see Fig. 1). Its complexity stems from the computation of K terminated random experiments with expected complexity $\mathcal{O}(np)$ (see Appendix E in the supplementary materials).
2. As inputs, the *T-Rex* selector requires only the data and the target FDR level. The tuning of the sparsity parameter for *Lasso*-type methods [37–40] is no longer required when incorporating them into the *T-Rex* selector framework.

In summary the *T-Rex* selector is, to the best of our knowledge, the first multivariate high-dimensional FDR-controlling method that scales to millions of variables in a reasonable amount of computation time (see Fig. 1), which makes it a suitable method for large-scale GWAS, i.e., our major use-case.

For other use-cases (e.g., high-dimensional survival analysis, sparse financial index tracking), where strong dependencies among the variables (e.g., gene expression levels, stock returns) exist and common SNP pruning or other preprocessing techniques are not applicable, the dependency-aware *T-Rex* (*T-Rex+DA*) selector has been proposed [41]. The *T-Rex+DA* selector performs high-dimensional FDR-controlled variable selection in the presence of strong dependencies at the cost of a reduced power compared to the *T-Rex* selector.

Moreover, the *T-Rex* selector has already proven to be a useful and versatile framework for screening large-scale genomics biobanks [42], efficient computation in big data applications [43], grouped variable selection [44,45], Gaussian graphical models [46], sparse principal component analysis [47], sparse financial index tracking [48], and survival analysis [41].

The open source R software packages *TRexSelector* [49] and *tlars* [50] contain the implementation of the proposed *T-Rex* selector.

Notation: Column vectors and matrices are denoted by boldface lowercase and uppercase letters, respectively. Scalars are denoted by non-boldface lowercase or uppercase letters. With the exceptions of $\mathcal{N}$ and $\varnothing$, which stand for the normal distribution and the empty set, respectively, sets are denoted by calligraphic uppercase letters, e.g., $\mathcal{A}$ with $|\mathcal{A}|$ denoting the associated cardinality. The symbols $\mathbb{E}$ and Var denote the expectation and the variance operator, respectively.

Organization: The remainder of this paper is organized as follows: Section 2 introduces the methodology of the proposed *T-Rex* selector. Section 3 presents the main theoretical results regarding the properties of the proposed method and its algorithmic details. Section 4 discusses the results of numerical simulations while Section 5 evaluates the performances of the proposed *T-Rex* selector and the benchmark methods on a simulated genome-wide association study (GWAS). Section 6 concludes the paper. Technical proofs, numerical verifications, additional simulations, and other appendices are deferred to the supplementary materials.

2. The T-Rex selector

This section introduces the proposed *T-Rex* selector. First, some forward variable selection methods that are used within the *T-Rex* selector are briefly revisited and a mathematical formulation of the FDR and TPR is given. Then, the underlying methodology is described and the optimization problem of calibrating the *T-Rex* selector to perform FDR control at the target level is formulated.

2.1. High-dimensional variable selection methods

The *T-Rex* selector framework is versatile in the sense that it can incorporate many different forward selection algorithms. In this paper, we will focus on *Lasso*-type methods [37,39,40,51] and especially the *LARS* algorithm [38]. Although, in general, the FDR control proof of the T-Rex selector (see Section 3.1) does not assume a linear relationship between the explanatory variables and the response variable, we will introduce the linear regression model because it is required by the high-dimensional forward variable selection methods that are considered in this paper.

The linear regression model is defined by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

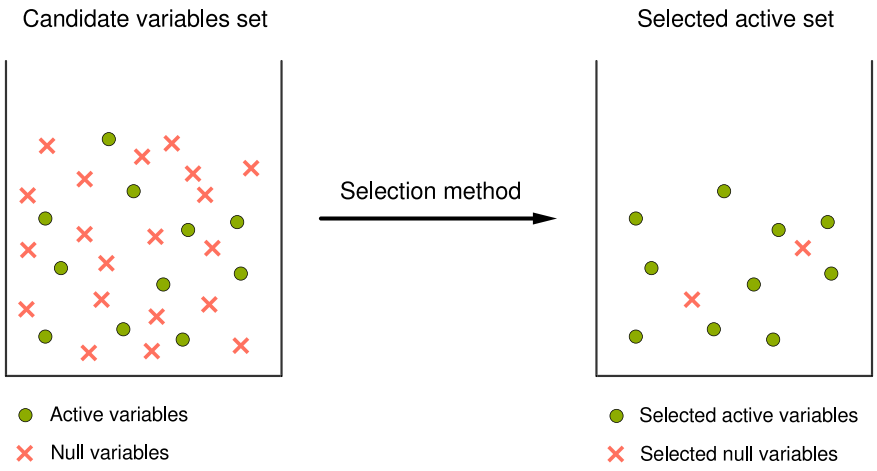


Fig. 2. Illustration of the general concept of active variables, selected active variables, null variables, and selected null variables. Note that active variables and null variables are also commonly referred to as true active variables and true null variables, respectively, to clearly distinguish them from the selected active and the selected null variables.

where $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \dots\ \boldsymbol{x}_p]$ with $\boldsymbol{x}_j \in \mathbb{R}^n$, $j = 1, \dots, p$, is the fixed predictor matrix containing p predictors and n observations, $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the parameter vector, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, with $\boldsymbol{I}$ being the identity matrix, is an additive Gaussian noise vector with standard deviation σ. Variables whose associated coefficients in $\boldsymbol{\beta}$ are non-zero (zero) are called actives or active variables (nulls or null variables).

To obtain a sparse estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, sparsity-inducing methods, such as the *Lasso* [37] and related methods [38–40,51], can be used. The *Lasso* solution is defined by

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\lambda > 0$ is a tuning parameter that controls the sparsity of $\hat{\boldsymbol{\beta}}$. Throughout this paper, we will use the closely related *LARS* algorithm [38] as a forward variable selection method to conduct the random experiments of the *T-Rex* selector. The solution path of the *Lasso* over λ is efficiently computed by applying a slightly modified *LARS* algorithm.[1] That is, instead of adding one variable at a time based on the highest correlation with the current residual, the *Lasso* modification requires the removal of previously added variables when the associated coefficients change their sign [38]. However, removed variables can enter the solution path again in later steps. Since the solution paths are terminated early by the *T-Rex* selector, there are very few or no zero crossings along these paths. Thus, in most cases, the *Lasso* in (2) and the *LARS* algorithm produce very similar or identical solution paths when used with the *T-Rex* selector.

2.2. FDR and TPR

Before providing a mathematical formulation of the FDR and the TPR, we first illustrate the concepts of 'active variables', 'selected active variables', 'null variables', 'selected null variables'. Consider a GWAS with SNPs as predictors and a disease of interest as the response variable Y. Active variables are the SNPs truly associated with the disease, while null variables are the SNPs not associated with the disease. Selected active variables are the associated SNPs chosen by the variable selection method, and selected null variables are the unassociated SNPs chosen by the variable selection method. For example, if SNPs X_1 and X_2 are truly associated with the disease and SNPs X_3 and X_4 are not, selecting X_1 and X_3 means X_1 is a selected active variable and X_3 is a selected null variable. The general concept of active variables, selected active variables, null variables, and selected null variables is illustrated in Fig. 2.

The FDR and TPR are expressed mathematically as follows: Given the index set of the active variables $\mathcal{A} \subseteq \{1, \dots, p\}$, where p is the number of candidate variables, and the index set of the selected active variables $\widehat{\mathcal{A}} \subseteq \{1, \dots, p\}$, the FDR and the TPR are defined by

$$\text{FDR} := \mathbb{E}\left[\frac{|\widehat{\mathcal{A}} \setminus \mathcal{A}|}{1 \vee |\widehat{\mathcal{A}}|}\right] \quad \text{and} \quad \text{TPR} := \mathbb{E}\left[\frac{|\mathcal{A} \cap \widehat{\mathcal{A}}|}{1 \vee |\mathcal{A}|}\right], \tag{3}$$

respectively, where $|\cdot|$ denotes the cardinality operator and the symbol $\vee$ stands for the maximum operator, i.e., $a \vee b = \max\{a, b\}$, $a, b \in \mathbb{R}$.[2]

Note that, throughout this work, p_1 and p_0 denote the number of active variables and the number of null variables, respectively. Thus, the number of candidate variables p is the sum of p_1 and p_0. Also, note that by definition the FDR and the TPR are zero when $|\widehat{\mathcal{A}}| = 0$ and $p_1 := |\mathcal{A}| = 0$, respectively. While the FDR and the TPR of an oracle variable selection procedure are 0% and 100%, respectively, in practice, a tradeoff must be accomplished.

2.3. The T-Rex selector: Methodology

The general methodology underpinning the *T-Rex* selector consists of several steps that are illustrated in Fig. 3. In the following, we will introduce the framework and the notation, which will be crucial for understanding why the *T-Rex* selector efficiently controls the FDR at the target level:

Step1: Generate $K > 1$ dummy matrices $\mathring{\boldsymbol{X}}_k$, $k = 1, \dots, K$, each containing $L \geq 1$ dummy predictors that are sampled from a standard normal distribution.

Step2: Append each dummy matrix to the original predictor matrix $\boldsymbol{X}$, resulting in the enlarged predictor matrices

$$\begin{aligned}\widetilde{\boldsymbol{X}}_k &:= [\boldsymbol{X}\ \mathring{\boldsymbol{X}}_k] \\ &= [\boldsymbol{x}_1\ \dots\ \boldsymbol{x}_p\ \mathring{\boldsymbol{x}}_{k,1}\ \dots\ \mathring{\boldsymbol{x}}_{k,L}], \quad k = 1, \dots, K,\end{aligned}$$

where $\mathring{\boldsymbol{x}}_{k,1}, \dots, \mathring{\boldsymbol{x}}_{k,L}$ are the dummies (see Fig. 4).

[1] An alternative approach to obtain the solution path of the *Lasso* is to apply the pathwise coordinate descent algorithm [52]. However, this approach is not a forward variable selection method and, therefore, not applicable within the *T-Rex* selector framework.

[2] Throughout this paper, the original definition of the FDR in [17] is used. Other definitions of the FDR, such as the positive FDR [53], exist. The interested reader is referred to both papers for discussions on different potential definitions of the FDR.

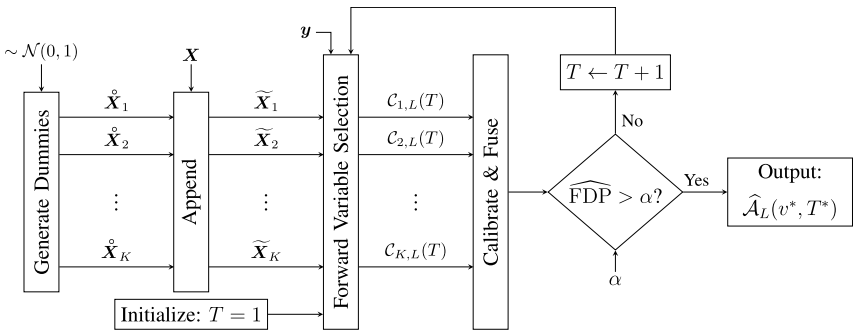


Fig. 3. Simplified overview of the *T-Rex* selector framework: For each random experiment $k \in \{1, \dots, K\}$, the *T-Rex* selector generates a dummy matrix $\mathring{\boldsymbol{X}}_k$ containing L dummies and appends it to $\boldsymbol{X}$ to obtain the enlarged predictor matrix $\widetilde{\boldsymbol{X}}_k = [\boldsymbol{X}\ \mathring{\boldsymbol{X}}_k]$. With $\widetilde{\boldsymbol{X}}_k$ and the response $\boldsymbol{y}$ as inputs, a forward variable selection method is applied to obtain the candidate sets $\mathcal{C}_{1,L}(T), \dots, \mathcal{C}_{K,L}(T)$, where T is iteratively increased from one until $\widehat{\text{FDP}}$ (i.e., an estimate of the proportion of false discoveries among all selected variables that is determined by the calibration process) exceeds the target FDR level $\alpha \in [0, 1]$. Finally, a fusion procedure determines the selected active set $\widehat{\mathcal{A}}_L(v^*, T^*)$ for which the calibration procedure provides the optimal parameters v^* and T^*, such that the FDR is controlled at the target level α while maximizing the number of selected variables.

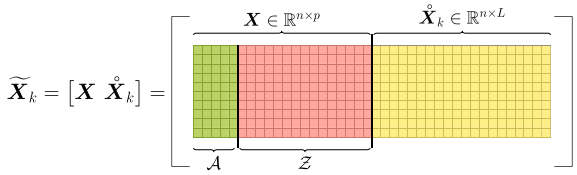


Fig. 4. The enlarged predictor matrices $\widetilde{\boldsymbol{X}}_k$, $k = 1, \dots, K$, replace the original predictor matrix $\boldsymbol{X}$ in each random experiment within the *T-Rex* selector framework. They contain the original and the dummy predictors. The index set of the active variables and the index set of the null variables are denoted by $\mathcal{A}$ and $\mathcal{Z}$, respectively. The number of active variables and the number of null variables are denoted by $p_1 := |\mathcal{A}|$ and $p_0 := |\mathcal{Z}|$, respectively.

Step3: Apply a forward variable selection procedure to $\{\widetilde{\boldsymbol{X}}_k, \boldsymbol{y}\}$, $k = 1, \dots, K$. For each random experiment, terminate the forward selection process after $T \geq 1$ dummy variables are included. This results in the candidate active sets $\mathcal{C}_{k,L}(T)$, $k = 1, \dots, K$. After terminating the forward selection process remove all dummies from the candidate active sets.[3]

Step4: Iteratively increase T and carry out Step3: until $\widehat{\text{FDP}}$ exceeds the target FDR level $\alpha \in [0, 1]$. $\widehat{\text{FDP}}$ is a conservative estimator of the false discovery proportion (FDP), i.e., the proportion of false discoveries among all selected variables. The relationship between the FDP and the FDR is that the FDR is the expected value of the FDP, i.e., FDR = $\mathbb{E}$[FDP]. FDR control is achieved by calibrating the *T-Rex* selector such that $\widehat{\text{FDP}}$ does not exceed the target FDR level α. Therefore, as depicted in Fig. 3, we cannot increase T further once $\widehat{\text{FDP}}$ exceeds α. The mathematical expression and the details of $\widehat{\text{FDP}}$ are deferred to Section 3. Also, the calibration process for determining $\widehat{\text{FDP}}$ and the optimal values v^* and T^* such that the FDR is controlled at the target level $\alpha \in [0, 1]$ while maximizing the number of selected variables is derived in Section 3.

Step5: Fuse the candidate active sets to determine the estimate of the active set $\widehat{\mathcal{A}}_L(v^*, T^*)$. The fusion step is based on the *relative occurrence* of the original variables:

Definition 1 (*Relative Occurrence*)**.** Let $K \in \mathbb{N}_+ \setminus \{1\}$ be the number of random experiments, $L \in \mathbb{N}_+$ the number of dummies, and $T \in \{1, \dots, L\}$ the number of included dummies after which the forward variable selection process in each random experiment is terminated. The relative occurrence of variable $j \in \{1, \dots, p\}$ is defined by

$$\boldsymbol{\Phi}_{T,L}(j) := \begin{cases} \dfrac{1}{K} \displaystyle\sum_{k=1}^{K} \mathbb{1}_k(j, T, L), & T \geq 1 \\ 0, & T = 0, \end{cases}$$

where $\mathbb{1}_k(j, T, L)$ is the indicator function for which

$$\mathbb{1}_k(j, T, L) = \begin{cases} 1, & j \in \mathcal{C}_{k,L}(T) \\ 0, & \text{otherwise.} \end{cases}$$

All variables whose relative occurrences at $T = T^*$ exceed the voting level $v^* \in [0.5, 1)$ are selected and the estimator of the active set is defined by

$$\widehat{\mathcal{A}}_L(v^*, T^*) := \{j : \boldsymbol{\Phi}_{T^*,L}(j) > v^*\}. \tag{4}$$

The details of how the calibration process determines T^* and v^* such that, for any choice of L, the *T-Rex* selector controls the FDR at the target level while maximizing the number of selected variables are deferred to Section 3.3. Moreover, an extension to the calibration process to jointly determine T^*, v^*, and L is also proposed in Section 3.3. The number of random experiments K is not subject to optimization. However, choosing $K \geq 20$ provides excellent empirical results and we never observed notable improvements for $K \geq 100$.[4]

An example that helps in developing an intuition for the three main ingredients of the *T-Rex* selector, which are (i) sampling dummies from a univariate distribution, (ii) early terminating the solution paths of

[3] Since we use the *LARS* method throughout this paper, variables can only be included but not dropped along the solution paths. Nevertheless, the *T-Rex* selector can also incorporate forward selection methods that remove some previously included variables from the candidate set along the solution path (e.g., *Lasso*). For such methods, the number of currently active dummies can decrease along the solution path. However, because the solution paths are terminated after T dummies are included for the first time, there is no ambiguity regarding the step in which the forward selection process ends.

[4] Instead of fixing the number of random experiments, it could be increased until the relative occurrences $\boldsymbol{\Phi}_{T,L}(j)$, $j = 1, \dots, p$, converge. However, since a significant reduction of computation time is achieved by executing the independent random experiments in parallel on multicore computers or high-performance clusters, fixing K to a multiple of the number of available CPUs is preferable.

the random experiments, and (iii) fusing the candidate sets based on relative occurrences, is deferred to Appendix B in the supplementary materials.

2.4. Problem statement

We formulate an optimization problem that maximizes the number of selected true positives and simultaneously controls the FDR at the target level. We start with some remarks on notation followed by definitions of the FDR and the TPR, which particularize the generic definitions in (3) for the *T-Rex* selector. For better readability, from now on, the arguments T and L of the estimator of the active set are dropped, i.e., $\widehat{\mathcal{A}}(\upsilon) := \widehat{\mathcal{A}}_L(\upsilon, T)$, except when referring specifically to the set in (4) for which the values υ^* and T^* result from the calibration that will be discussed in Section 3. Note that "included candidates" refers to the variables picked (and not dropped) along the solution path of each random experiment. "Selected variables" refers to the variables whose relative occurrences exceed the voting level $\upsilon \in [0.5, 1)$.

Definition 2 ($V_{T,L}(\upsilon)$, $S_{T,L}(\upsilon)$ *and* $R_{T,L}(\upsilon)$)**.** The number of selected null variables $V_{T,L}(\upsilon)$, the number of selected active variables $S_{T,L}(\upsilon)$, and the number of selected variables $R_{T,L}(\upsilon)$ are defined, respectively, by

$$V_{T,L}(\upsilon) := \left|\widehat{\mathcal{A}}^0(\upsilon)\right| := \left|\{\text{null } j : \Phi_{T,L}(j) > \upsilon\}\right|,$$
$$S_{T,L}(\upsilon) := \left|\widehat{\mathcal{A}}^1(\upsilon)\right| := \left|\{\text{active } j : \Phi_{T,L}(j) > \upsilon\}\right|, \text{ and}$$
$$R_{T,L}(\upsilon) := V_{T,L}(\upsilon) + S_{T,L}(\upsilon) = \left|\widehat{\mathcal{A}}(\upsilon)\right|.$$

The FDR and TPR expressions in (3) are rewritten using Definition 2 as follows:

Definition 3 (*FDP and FDR*)**.** The false discovery proportion (FDP) is defined by

$$\mathrm{FDP}(\upsilon, T, L) := \frac{V_{T,L}(\upsilon)}{R_{T,L}(\upsilon) \vee 1}$$

and the FDR is defined by

$$\mathrm{FDR}(\upsilon, T, L) := \mathbb{E}\big[\mathrm{FDP}(\upsilon, T, L)\big],$$

where the expectation is taken with respect to the noise in (1).

Definition 4 (*TPP and TPR*)**.** The true positive proportion (TPP) is defined by

$$\mathrm{TPP}(\upsilon, T, L) := \frac{S_{T,L}(\upsilon)}{p_1 \vee 1}$$

and the TPR is defined by

$$\mathrm{TPR}(\upsilon, T, L) := \mathbb{E}\big[\mathrm{TPP}(\upsilon, T, L)\big],$$

where the expectation is taken with respect to the noise in (1).

Remark 1. Note that if $R_{T,L}(\upsilon)$ is equal to zero, then $V_{T,L}(\upsilon)$ is zero as well. In this case, the denominator in the expression for the FDP is set to one and, thus, the FDP becomes zero. This is a reasonable solution to the "0/0" case, because when no variables are selected there exist no false discoveries. Similarly, when there exist no true active variables among the candidates, i.e. $p_1 = S_{T,L}(\upsilon) = 0$, the TPP equals zero.

A major result of this work is to determine T^* and υ^*, such that, for any fixed $L \in \mathbb{N}_+$, the *T-Rex* selector maximizes $\mathrm{TPR}(\upsilon, T, L)$ while provably controlling $\mathrm{FDR}(\upsilon, T, L)$ at any given target level $\alpha \in [0, 1]$. In practice, this amounts to finding the solution of the optimization problem

$$\max_{\upsilon, T} \mathrm{TPP}(\upsilon, T, L) \quad \text{s.t.} \quad \widehat{\mathrm{FDP}}(\upsilon, T, L) \le \alpha, \tag{5}$$

which is equivalent to

$$\max_{\upsilon, T} S_{T,L}(\upsilon) \quad \text{s.t.} \quad \widehat{\mathrm{FDP}}(\upsilon, T, L) \le \alpha \tag{6}$$

because p_1 is a constant. Note that $\widehat{\mathrm{FDP}}(\upsilon, T, L)$ is a conservative estimator of $\mathrm{FDP}(\upsilon, T, L)$, i.e., it holds that $\mathrm{FDR}(\upsilon, T, L) = \mathbb{E}\big[\mathrm{FDP}(\upsilon, T, L)\big] \le \mathbb{E}\big[\widehat{\mathrm{FDP}}(\upsilon, T, L)\big] = \widehat{\mathrm{FDR}}(\upsilon, T, L)$. The details of the conservative FDP estimator are discussed in Section 3. Since the number of true active variables, $S_{T,L}(\upsilon)$, is unobservable, it is standard to use $R_{T,L}(\upsilon)$ (i.e., the observable total number of selected variables) as a practical surrogate. This approach is adopted by all FDR-controlling benchmark methods mentioned in Section 1. This results in the final optimization problem:

$$\max_{\upsilon, T} R_{T,L}(\upsilon) \quad \text{s.t.} \quad \widehat{\mathrm{FDP}}(\upsilon, T, L) \le \alpha. \tag{7}$$

In words: *The T-Rex selector maximizes the number of selected variables while controlling a conservative estimator of the FDP at the target level* α.

Note that the reason why $R_{T,L}(\upsilon)$ is a practical surrogate for $S_{T,L}(\upsilon)$ in the context of FDR control is that it is (i) an observable upper bound of $S_{T,L}(\upsilon)$ and (ii) a good approximation of $S_{T,L}(\upsilon)$ for reasonably low target FDR levels α. Mathematically, this relationship is obtained by rewriting the FDR in Definition 3 using Definition 2, which yields

$$\mathrm{FDR}(\upsilon, T, L) = \mathbb{E}\left[\frac{R_{T,L}(\upsilon) - S_{T,L}(\upsilon)}{R_{T,L}(\upsilon) \vee 1}\right] \le \alpha.$$

That is, the relative difference of $R_{T,L}(\upsilon)$ and $S_{T,L}(\upsilon)$ is upper-bounded by α on average. Thus, for small user-defined target FDR levels α, $R_{T,L}(\upsilon)$ is a good approximation of $S_{T,L}(\upsilon)$. For these two reasons, (7) has become the standard optimization problem throughout the FDR control literature and especially, as mentioned above, has been adopted by all benchmark methods.

In Section 3, it is shown that the T-Rex selector efficiently solves (7). Note that since (5), (6), and (7) share the same feasible region, any solution of (7) is trivially a feasible solution of (5) and (6).

3. Main results

This section contains our main results about the proposed *T-Rex* selector, which concern: FDR-control (Theorem 1), dummy generation (Theorem 2), and the optimal calibration algorithm (Theorem 3). We use martingale theory [36] to prove the FDR control property of the *T-Rex* selector. The developed FDR control theory relies on standard assumptions that are extensively verified especially for GWAS, i.e., the main use-case of this paper (see Appendices F, G, and J in the supplementary materials). Additionally, the computational complexity of the *T-Rex* selector, which stems from the computation of K terminated random experiments with expected complexity $\mathcal{O}(np)$, is derived in Appendix E in the supplementary materials.

3.1. FDR control

In Definition 1, the relative occurrence $\Phi_{T,L}(j)$ of the jth candidate variable has been introduced. It can be decomposed into the changes in relative occurrence, i.e.,

$$\Phi_{T,L}(j) = \sum_{t=1}^{T} \Delta\Phi_{t,L}(j), \quad j = 1, \dots, p,$$

where $\Delta\Phi_{t,L}(j) := \Phi_{t,L}(j) - \Phi_{t-1,L}(j)$ is the change in relative occurrence from step $t-1$ to t for variable j.[5]

[5] When using a forward selection method within the *T-Rex* selector framework that does not drop variables along the solution path (e.g. *LARS*), all $\Phi_{t,L}(j)$'s are non-decreasing in t and, therefore, $\Delta\Phi_{t,L}(j) \ge 0$ for all j. In contrast, when using forward selection methods that might drop variables along the solution path (e.g. *Lasso*), the $\Phi_{t,L}(j)$'s might decrease in t and, therefore, the $\Delta\Phi_{t,L}(j)$'s can be negative. Nevertheless, the relative occurrence $\Phi_{T,L}(j)$ is non-negative for all j and any forward selection method.

Since the active and the null variables are interspersed in the solution paths of the random experiments, some null variables might appear earlier on the solution paths than some active variables. Many researchers have observed that active and null variables are interspersed in solution paths obtained from sparsity-inducing methods, such as the *LARS* algorithm or the *Lasso* [19,54]. Therefore, it is unavoidable that the $\Delta\Phi_{t,L}(j)$'s of the null variables are inflated, meaning that their values become artificially larger than expected, along the solution paths of the random experiments. Moreover, we observe interspersion not only for active and null variables but also for dummies, which is expected since dummies can be interpreted as flagged null variables.

The above considerations motivate the definition of the *deflated relative occurrence* to harness the information about the fraction of included dummies in each step along the solution paths in order to deflate the $\Delta\Phi_{t,L}(j)$'s of the null variables and, thus, account for the interspersion effect.

Definition 5 (*Deflated relative occurrence*)**.** The deflated relative occurrence of variable j is defined by

$$\Phi'_{T,L}(j) := \sum_{t=1}^{T}\left(1 - \frac{p - \sum_{q=1}^{p}\Phi_{t,L}(q)}{L-(t-1)}\frac{1}{\sum_{q\in\widehat{\mathcal{A}}(0.5)}\Delta\Phi_{t,L}(q)}\right)\Delta\Phi_{t,L}(j),$$

$j = 1,\dots,p$.

In words: *The deflated relative occurrence is the sum over the deflated $\Delta\Phi_{t,L}(j)$'s from step $t=1$ until step $t=T$*. As detailed and intuitively explained in Appendix C in the supplementary materials, the $\Delta\Phi_{t,L}(j)$'s are multiplied by a deflation factor that takes into account the ratio between the fraction of selected dummies and the fraction of selected candidate variables in each step $t \in \{1,\dots,T\}$.

The reader might wonder whether the deflation factors affect the $\Delta\Phi_{t,L}(j)$'s of active variables in addition to those of null variables. Deflation factors minimally impact the $\Delta\Phi_{t,L}(j)$'s of active variables because active variables typically enter the solution paths early at low values of t, where the deflation factor is close to one, and at higher t, $\Delta\Phi_{t,L}(j)$ for active variables approaches zero. A more detailed explanation of this is deferred to Appendix C in the supplementary materials.

Using the deflated relative occurrences, the estimator of $V_{T,L}(\upsilon)$, i.e., the number of selected null variables (see Definition 2), and the corresponding FDP estimator are defined as follows:

Definition 6 (*FDP estimator*)**.** The estimator of $V_{T,L}(\upsilon)$ is defined by

$$\widehat{V}_{T,L}(\upsilon) := \sum_{j\in\widehat{\mathcal{A}}(\upsilon)}\left(1-\Phi'_{T,L}(j)\right)$$

and the corresponding estimator of $\mathrm{FDP}(\upsilon,T,L)$ is defined by

$$\widehat{\mathrm{FDP}}(\upsilon,T,L) = \frac{\widehat{V}_{T,L}(\upsilon)}{R_{T,L}(\upsilon)\vee 1} \tag{8}$$

with

$$\widehat{\mathrm{FDR}}(\upsilon,T,L) := \mathbb{E}\left[\widehat{\mathrm{FDP}}(\upsilon,T,L)\right]$$

being its expected value.

Note that the term $\widehat{V}_{T,L}(\upsilon)$ is an estimator of the number of false discoveries within the set of selected variables $\widehat{\mathcal{A}}(\upsilon)$. This estimator is obtained by summing up the complements of the deflated relative occurrences $\Phi'_{T,L}(j)$ for each selected variable j. Essentially, $1-\Phi'_{T,L}(j)$ estimates the probability that a selected variable j is a false discovery, and summing these values across all selected variables provides the total estimate of false discoveries.

The main idea behind FDR control for the *T-Rex* selector is that controlling $\widehat{\mathrm{FDP}}(\upsilon,T,L)$ at the target level $\alpha \in [0,1]$ guarantees that $\mathrm{FDR}(\upsilon,T,L)$ is controlled at the target level as well. To achieve this, we define $\upsilon \in [0.5,1)$ as the voting level at which $\widehat{\mathrm{FDP}}(\upsilon,T,L)$ is controlled at the target level. Note that υ has to be at least 50% to ensure that all selected variables occur in at least more than the majority of the candidate sets within the *T-Rex* selector.

Definition 7 (*Voting level*)**.** Let $T \in \{1,\dots,L\}$ and $L \in \mathbb{N}_+$ be fixed. Then, the voting level is defined by

$$\upsilon := \inf\{\nu \in [0.5,1) : \widehat{\mathrm{FDP}}(\nu,T,L) \le \alpha\} \tag{9}$$

with the convention that $\upsilon = 1$ if the infimum does not exist.[6]

Remark 2. Recall that the aim that is stated in the optimization problem in (7) is to select as many variables as possible while controlling $\widehat{\mathrm{FDP}}(\upsilon,T,L)$ at the target level. For fixed T and L, this is achieved by the smallest voting level that satisfies the constraint on $\widehat{\mathrm{FDP}}(\upsilon,T,L)$. We can easily see that for any fixed T and L, the voting level in (9) solves the optimization problem in (7). The reason is that for any two voting levels $\upsilon_1, \upsilon_2 \in [0.5,1)$ with $\upsilon_2 \ge \upsilon_1$ satisfying the $\widehat{\mathrm{FDP}}$-constraint in (9), it holds that $R_{T,L}(\upsilon_1) \ge R_{T,L}(\upsilon_2)$.

Remark 3. If υ, T, and L satisfy Eq. (9), then the FDP from Definition 3 can be upper-bounded as follows:

$$\begin{aligned}\mathrm{FDP}(\upsilon,T,L) &= \frac{V_{T,L}(\upsilon)}{R_{T,L}(\upsilon)\vee 1} = \widehat{\mathrm{FDP}}(\upsilon,T,L)\cdot\frac{V_{T,L}(\upsilon)}{\widehat{V}_{T,L}(\upsilon)}\\ &\le \alpha\cdot\frac{V_{T,L}(\upsilon)}{\widehat{V}_{T,L}(\upsilon)} \le \alpha\cdot\frac{V_{T,L}(\upsilon)}{\widehat{V}'_{T,L}(\upsilon)},\end{aligned}$$

where $\widehat{V}'_{T,L}(\upsilon)$, which is supposed to be greater than zero, is defined by

$$\widehat{V}'_{T,L}(\upsilon) := \widehat{V}_{T,L}(\upsilon) - \sum_{j\in\widehat{\mathcal{A}}(\upsilon)}\left(1-\Phi_{T,L}(j)\right).$$

Before the FDR control theorem is formulated, we introduce a lemma that contains the backbone of our FDR control theorem, which is rooted in martingale theory [36]:

Lemma 5. *Define $\mathcal{V} := \{\Phi_{T,L}(j) : \Phi_{T,L}(j) > 0.5,\ j=1,\dots,p\}$ and*

$$H_{T,L}(\upsilon) := \frac{V_{T,L}(\upsilon)}{\widehat{V}'_{T,L}(\upsilon)}.$$

Let $\mathcal{F}_\upsilon := \sigma\left(\{V_{T,L}(u)\}_{u\ge\upsilon}, \{\widehat{V}'_{T,L}(u)\}_{u\ge\upsilon}\right)$ be a backward-filtration with respect to υ. Then, for all tuples $(T,L) \in \{1,\dots,L\}\times\mathbb{N}_+$, $\{H_{T,L}(\upsilon)\}_{\upsilon\in\mathcal{V}}$ is a backward-running super-martingale with respect to $\mathcal{F}_\upsilon$. That is,

$$\mathbb{E}\left[H_{T,L}\left(\upsilon-\epsilon^*_{T,L}(\upsilon)\right)\,\middle|\,\mathcal{F}_\upsilon\right] \ge H_{T,L}(\upsilon),$$

where

$$\epsilon^*_{T,L}(\upsilon) := \inf\{\epsilon\in(0,\upsilon) : R_{T,L}(\upsilon-\epsilon) - R_{T,L}(\upsilon) = 1\}$$

with $\upsilon \in [0.5,1)$ and the convention that $\epsilon^*_{T,L}(\upsilon) = 0$ if the infimum does not exist.

Proof. The proof is deferred to Appendix A in the supplementary materials. □

Theorem 1 (*FDR control*)**.** *Suppose that $\widehat{V}'_{T,L}(\upsilon) > 0$. Then, for all triples $(T,L,\upsilon) \in \{1,\dots,L\}\times\mathbb{N}_+\times[0.5,1)$ that satisfy Eq. (9) and as $K\to\infty$, the T-Rex selector controls the FDR at any fixed target level $\alpha\in[0,1]$, i.e.,*

$$\mathrm{FDR}(\upsilon,T,L) = \mathbb{E}\left[\mathrm{FDP}(\upsilon,T,L)\right] \le \alpha.$$

[6] The voting level can be interpreted as a stopping time. The term 'stopping time' stems from martingale theory [36]. In the proof of Lemma 5 in Appendix A in the supplementary materials, it is shown that indeed υ is a stopping time with respect to some still to be defined filtration of a still to be defined stochastic process. Note that the convention of setting $\upsilon = 1$ if the infimum does not exist ensures that no variables are selected when there exists no triple (T,L,υ) that satisfies Eq. (9).

Proof sketch. Taking the expectation on both sides of the inequality in Remark 3 yields an upper bound on the FDR, i.e.,

$$\mathrm{FDR}(\upsilon, T, L) = \mathbb{E}\big[\mathrm{FDP}(\upsilon, T, L)\big] \leq \alpha \cdot \mathbb{E}\big[H_{T,L}(\upsilon)\big],$$

where $H_{T,L}(\upsilon) = V_{T,L}(\upsilon)/\widehat{V}'_{T,L}(\upsilon)$, as defined within Lemma 5. To prove FDR control at the target level α, it remains to prove that $\mathbb{E}\big[H_{T,L}(\upsilon)\big] \leq 1$. This is achieved by using the fact that the process $\{H_{T,L}(\upsilon)\}_{\upsilon \in \mathcal{V}}$ is a backward-running super-martingale, as stated in Lemma 5. This martingale property allows using the optional stopping theorem [36] to upper-bound the expected value of $H_{T,L}(\upsilon)$ by its expected value at the initial point $\upsilon = 0.5$, which is smaller than one, i.e., $\mathbb{E}\big[H_{T,L}(\upsilon)\big] \leq \mathbb{E}\big[H_{T,L}(0.5)\big] \leq 1$. □

The details of the proof are deferred to Appendix A in the supplementary materials. Along with the details of the proof of the FDR control theorem, detailed explanations of the standard assumptions and their extensive numerical verifications are deferred to Appendices F, G, and J in the supplementary materials.

3.2. Dummy generation

As shown in Fig. 3, the *T-Rex* selector generates L i.i.d. dummies for each random experiment by sampling each element of the dummy vectors from the standard normal distribution, i.e.,

$$\mathring{\boldsymbol{x}}_l = [\mathring{x}_{1,l} \ \dots \ \mathring{x}_{n,l}]^\top, \ \text{where} \ \mathring{x}_{i,l} \sim \mathcal{N}(0,1),$$

$i = 1, \dots, n$, $l = 1, \dots, L$. This raises the question whether dummies can be sampled from other distributions, as well, to serve as flagged null variables. From an asymptotic point of view, i.e., $n \to \infty$, and if some mild conditions are satisfied, the perhaps at first glance surprising answer to this question is that *dummies can be sampled from any univariate probability distribution with finite expectation and variance in order to serve as flagged null variables within the T-Rex selector*.

We will prove the above statement for any correlation-based forward selection procedure. Specifically, this includes procedures that use sample correlations of the predictors with the response or with the current residuals in each forward selection step to determine which variable is included next. Thus, the statement is true, e.g., for the *LARS* algorithm, *Lasso*, *adaptive Lasso*, and *elastic net*.

Recall that null variables and dummies are not related to the response. For null variables this holds by definition and for dummies this holds because dummies are generated without using any information about the response.[7] Moreover, the sample correlations of the dummies with the response are random. Thus, the higher the number of generated dummies, the higher the probability of including a dummy instead of a null or even a true active variable in the next step of a random experiment. These considerations suggest that only the number of dummies within the enlarged predictor matrices is relevant for the behavior of the forward selection process in each random experiment. That is, the core of the following Theorem 2 is that, for $n \to \infty$, the distribution from which the dummies are sampled has no influence on the distribution of the correlation variables

$$\mathring{G}_{l,m,k} := \sum_{i=1}^{n} \gamma_{i,m,k} \cdot \mathring{X}_{i,l,k},$$

$l \in \mathcal{D}_{m,k}$, $m \geq 1$, $k = 1, \dots, K$, where $\gamma_{i,m,k}$ is the ith element of $\boldsymbol{\gamma}_{m,k} := \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{m,k}$ (i.e., the residual vector in the mth forward selection step of the kth random experiment) with $\hat{\boldsymbol{\beta}}_{m,k}$ and $\mathcal{D}_{m,k}$ being the estimator of the parameter vector and the index set of the non-included dummies in the mth forward selection step of the kth random experiment, respectively. Note that $\boldsymbol{\gamma}_{1,k} = \boldsymbol{y}$ for all k, since $\hat{\boldsymbol{\beta}}_{1,k} = \boldsymbol{0}$ for all k, i.e., the residual vector in the first step of the forward selection process is simply the response vector $\boldsymbol{y}$. The random variable $\mathring{X}_{i,l,k}$ represents the ith element of the lth dummy within the kth random experiment. Summarizing, $\mathring{G}_{l,m,k}$ can be interpreted as the weighted sum of the i.i.d. random variables $\mathring{X}_{1,l,k}, \dots, \mathring{X}_{n,l,k}$ with fixed weights $\gamma_{1,m,k}, \dots, \gamma_{n,m,k}$. With these preliminaries in place, the second main theorem is formulated as follows:

Theorem 2 (*Dummy generation*). *Let $\mathring{X}_{i,l,k}$, $i = 1, \dots, n$, $l \in \mathcal{D}_{m,k}$, $m \geq 1$, $k = 1, \dots, K$, be standardized i.i.d. dummy random variables (i.e., $\mathbb{E}\big[\mathring{X}_{i,l,k}\big] = 0$ and $\mathrm{Var}\big[\mathring{X}_{i,l,k}\big] = 1$ for all i, l, m, k) following any probability distribution with finite expectation and variance. Define*

$$D_{n,l,m,k} := \frac{1}{\Gamma_{n,m,k}} \cdot \mathring{G}_{l,m,k},$$

where $\Gamma^2_{n,m,k} := \sum_{i=1}^{n} \gamma^2_{i,m,k}$ with $\Gamma_{n,m,k} > 0$ for all n, m, k and with fixed $\gamma_{i,m,k} \in \mathbb{R}$ for all i, m, k. Suppose that

$$\lim_{n \to \infty} \frac{\gamma_{i,m,k}}{\Gamma_{n,m,k}} = 0, \quad i = 1, \dots, n,$$

for all m, k. Then, as $n \to \infty$,

$$D_{n,l,m,k} \xrightarrow{d} D, \quad D \sim \mathcal{N}(0,1),$$

for all l, m, k.

Proof sketch. The Lindeberg–Feller central limit theorem is applicable because $\mathring{X}_{i,l,k}$, $i = 1, \dots, n$, $l \in \mathcal{D}_{m,k}$, $m \geq 1$, $k = 1, \dots, K$, are i.i.d. random variables and it holds that $\mathbb{E}\big[D_{n,l,m,k}\big] = 0$ and $\mathrm{Var}\big[D_{n,l,m,k}\big] = 1$. Moreover, since $\mathring{Q}_{i,l,m,k} := \gamma_{i,m,k} \cdot \mathring{X}_{i,l,k} \,/\, \Gamma_{n,m,k}$ satisfies the Lindeberg condition for all l, m, k, the theorem follows. □

The details of the proof and illustrative examples with non-Gaussian dummies are deferred to Appendix A and Appendix K, respectively, in the supplementary materials.

The reason why the theorem is an asymptotic result that requires $n \to \infty$ is that its proof uses the Lindeberg–Feller central limit theorem. Nevertheless, the widespread adoption of several flavors of the central limit theorem stems from the well-known fact that it usually provides already good convergence results at low sample sizes. Our additional simulations in Appendix K of the supplementary materials with dummies sampled not only from the standard normal distribution but also from non-Gaussian distributions (i.e., uniform, Student-t, and Gumbel) numerically verify Theorem 2 in a high-dimensional setting with $n = 300$ samples and $p = 1000$ variables. For all these dummy distributions, the results are almost identical. That is, there is no superior dummy distribution and, therefore, we simply choose to sample dummies from the standard normal distribution throughout this work.

Remark 4. Note that sampling dummies from any univariate probability distribution with finite expectation and variance to serve as flagged null variables is only reasonable in combination with multiple random experiments as conducted by the proposed *T-Rex* selector. We emphasize that Theorem 2 is not applicable to knockoff generation procedures of, e.g., *fixed-X* and *model-X* knockoffs.

3.3. The T-Rex selector: Optimal calibration algorithm

This section describes the proposed *T-Rex* calibration algorithm, which efficiently solves the optimization problem in (7) and provides feasible solutions for (5) and (6). The pseudocode of the *T-Rex* calibration method is provided in Algorithm 1. The algorithm flow is as follows: First, the number of dummies L and the number of random

[7] Note that the knockoff generation processes of the *fixed-X* and the *model-X* knockoff method, i.e., the benchmark methods, are fundamentally different from our approach that uses dummies. Although these methods also do not use any information about the response to generate the knockoffs, unlike the proposed *T-Rex* selector, they must incorporate the covariance structure of the predictor matrix, which leads to a large computation time, especially for high dimensions (see Appendix B in the supplementary materials and Figure 1).

experiments K are set (usually $L = p$ and $K = 20$).[8] Then, setting $\upsilon = 1 - \Delta\upsilon$ and starting at $T = 1$, the number of included dummies is iteratively increased until reaching the value of T for which the FDP estimate at a voting level of $\upsilon = 1 - \Delta\upsilon$ exceeds the target level for the first time. In each iteration, before the target level is exceeded, $\widehat{\mathcal{A}}_L(\upsilon, T)$ is computed as in (4) on a grid for υ, while for values of υ for which $\widehat{\mathrm{FDP}}(\upsilon, T, L)$ exceeds the target level $\widehat{\mathcal{A}}_L(\upsilon, T)$ is equal to the empty set. Picking the υ' and T' that maximize the number of selected variables yields the final solution.[9]

Algorithm 1 *T-Rex* Calibration.

1. **Input**: $\alpha \in [0, 1]$, K, L, $\boldsymbol{X}$, $\boldsymbol{y}$.
2. **Set** $T = 1$, $\Delta\upsilon = \frac{1}{K}$, $\widehat{\mathrm{FDP}}(\upsilon = 1 - \Delta\upsilon, T, L) = 0$.
3. **While** $\widehat{\mathrm{FDP}}(\upsilon = 1 - \Delta\upsilon, T, L) \leq \alpha$ and $T \leq L$ **do**:
   - 3.1. **For** $\upsilon = 0.5, 0.5 + \Delta\upsilon, 0.5 + 2 \cdot \Delta\upsilon, \ldots, 1 - \Delta\upsilon$ **do**:
     - i. **Compute** $\widehat{\mathrm{FDP}}(\upsilon, T, L)$ as in (8).
     - ii. **If** $\widehat{\mathrm{FDP}}(\upsilon, T, L) \leq \alpha$
       - **Compute** $\widehat{\mathcal{A}}_L(\upsilon, T)$ as in (4).
     - **Else**
       - **Set** $\widehat{\mathcal{A}}_L(\upsilon, T) = \varnothing$.
   - 3.2. **Set** $T \leftarrow T + 1$.
4. **Solve**

$$\max_{\upsilon', T'} \left|\widehat{\mathcal{A}}_L(\upsilon', T')\right|$$
$$\text{s.t. } T' \in \{1, \ldots, T - 1\}$$
$$\upsilon' \in \{0.5, 0.5 + \Delta\upsilon, 0.5 + 2 \cdot \Delta\upsilon, \ldots, 1 - \Delta\upsilon\}$$

   and let (υ^*, T^*) be a solution.
5. **Output**: (υ^*, T^*) and $\widehat{\mathcal{A}}_L(\upsilon^*, T^*)$.

The reason for exiting the loop in Step 3 when the FDP estimate at a voting level of $1 - \Delta\upsilon$ exceeds the target level for the first time is based on two key observations from our still to be presented simulation results (see Fig. 5):

1. For any fixed T and L the average value of $\widehat{\mathrm{FDP}}(\upsilon, T, L)$ decreases as υ increases.
2. For any fixed υ and L the average value of $\widehat{\mathrm{FDP}}(\upsilon, T, L)$ increases as T increases.

Remark 5. To foster the intuition behind these observations, we note that Eq. (8) can be written as follows:

$$\widehat{\mathrm{FDP}}(\upsilon, T, L) = \frac{\widehat{V}_{T,L}(\upsilon)}{\left(V_{T,L}(\upsilon) + S_{T,L}(\upsilon)\right) \vee 1}.$$

Taking Definitions 2, 6, and the reformulation of Eq. (8) into account, we see that the observations suggest that we can expect the rather conservative estimate $\widehat{V}_{T,L}(\upsilon)$ of $V_{T,L}(\upsilon)$ in the numerator to decrease faster than the total number of selected variables $V_{T,L}(\upsilon) + S_{T,L}(\upsilon)$ in the denominator when increasing the voting level υ. This is something that can be expected since, in general, assuming a variable selection method that performs better than random selection, active variables are expected to have higher relative occurrences than null variables and, therefore, remain selected even for large values of the voting level υ. A similar reasoning can be applied to intuitively understand the monotonical increase of $\widehat{\mathrm{FDP}}(\upsilon, T, L)$ with respect to T.

With these preliminaries in place, the third main theorem of this paper can be formulated:

Theorem 3 (*Optimality of Algorithm 1*). *Let (υ^*, T^*) be a solution determined by Algorithm 1 and suppose that, ceteris paribus, $\widehat{\mathrm{FDP}}(\upsilon, T, L)$ is monotonically decreasing in υ and monotonically increasing in T. Then, (υ^*, T^*) is an optimal solution of (7) and a feasible solution of (5) and (6).*

Proof sketch. Since the objective functions of the optimization problems in Step 4 of Algorithm 1 and in (7) are equivalent, i.e., $\left|\widehat{\mathcal{A}}_L(\upsilon, T)\right| = R_{T,L}(\upsilon)$, it only needs to be shown that the feasible set in Step 4 of the algorithm contains the feasible set of (7). Since the conditions of the optimization problems in (5), (6), and (7) are equivalent, this also proves that (υ^*, T^*) is a feasible solution of (5) and (6). □

The details of the proof are deferred to Appendix A in the supplementary materials.

3.4. Extension to the calibration algorithm

In Theorem 1, we have also established that the *T-Rex* selector controls the FDR at the target level for any choice of the number of dummies L. However, the choice of L has an influence on how tightly the FDR is controlled at the target level (see Fig. 5). Since controlling the FDR more tightly usually increases the TPR (i.e., power), it is desirable to choose the parameters of the *T-Rex* selector accordingly. We will see in the simulations in Section 4 that with increasing L, the FDR can be more tightly controlled at low target levels. In order to harness the positive effects that come with larger values of L while limiting the increased memory requirement for high values of L, we propose an extended version of the calibration algorithm that jointly determines υ, T, and L such that the FDR is more tightly controlled at the target FDR level while not running out of memory.[10] The major difference to Algorithm 1 is that the number of dummies L is iteratively increased until the estimate of the FDP falls below the target FDR level α. The pseudocode of the extended *T-Rex* calibration algorithm is provided in Algorithm 2.[11]

Note that the extension to Algorithm 1 lies in Step 2 and Step 3. Additionally, and in contrast to Algorithm 1, the input to the algorithm is extended by a reference voting level $\tilde{\upsilon} \in [0.5, 1)$ and the maximum values of L and T, namely $L_{\max}$ and $T_{\max}$. The algorithm flow is as follows: First L and T are set as follows: $L = p$ and $T = 1$. Then, starting at $L = p$ the number of dummies L is iteratively increased in steps of p until the estimate of the FDP at the voting level $\tilde{\upsilon}$ falls below the

[8] As already mentioned in Section 2.3, K is not subject to optimization. In practice, choosing $K = 20$ already provides excellent results (see Section 4) and only incremental improvements are achieved with larger values of K.

[9] In case of multiple solutions, we recommend to choose the solution with the largest υ because such a solution provides the variables that were selected most frequently. Nevertheless, all solutions to the calibration problem that are computed using Algorithm 1 provide FDR control while maximizing the number of selected variables.

[10] The reader might raise the question whether also the computation time increases with increasing L. There is no definite answer to this question. On the one hand, for very large values of L the computation time might increase. On the other hand, with increasing L the solution paths of the experiments are terminated earlier because the probability of selecting dummies grows with increasing L. Thus, increasing L might increase or decrease the computation time depending on whether the first or the second effect dominates.

[11] The R package *TRexSelector* [49] contains the implementation of the extended calibration algorithm in Algorithm 2.

Algorithm 2 Extended *T-Rex* Calibration.

1. **Input**: $\alpha \in [0,1]$, K, $\boldsymbol{X}$, $\boldsymbol{y}$, $\tilde{v}$, $L_{\max}$, $T_{\max}$.
2. **Set** $L = p$, $T = 1$.
3. **While** $\widehat{\mathrm{FDP}}(v = \tilde{v}, T, L) > \alpha$ and $L \leq L_{\max}$ **do**:
   Set $L \leftarrow L + p$.
4. **Set** $\Delta v = \frac{1}{K}$, $\widehat{\mathrm{FDP}}(v = 1 - \Delta v, T, L) = 0$.
5. **While** $\widehat{\mathrm{FDP}}(v = 1 - \Delta v, T, L) \leq \alpha$ and $T \leq T_{\max}$ **do**:
   5.1. **For** $v = 0.5, 0.5 + \Delta v, 0.5 + 2 \cdot \Delta v, \dots, 1 - \Delta v$ **do**:
      i. **Compute** $\widehat{\mathrm{FDP}}(v, T, L)$ as in (8).
      ii. **If** $\widehat{\mathrm{FDP}}(v, T, L) \leq \alpha$
         Compute $\widehat{\mathcal{A}}_L(v, T)$ as in (4).
      Else
         Set $\widehat{\mathcal{A}}_L(v, T) = \emptyset$.
   5.2. **Set** $T \leftarrow T + 1$.
6. **Solve**

$$\max_{v', T'} \left| \widehat{\mathcal{A}}_L(v', T') \right|$$
$$\text{s.t. } T' \in \{1, \dots, T-1\}$$
$$v' \in \{0.5, 0.5 + \Delta v, 0.5 + 2 \cdot \Delta v, \dots, 1 - \Delta v\}$$

   and let (v^*, T^*) be a solution.
7. **Output**: (v^*, T^*) and $\widehat{\mathcal{A}}_L(v^*, T^*)$.

target FDR level α or L exceeds $L_{\max}$. The rest of the algorithm is as in Algorithm 1 except that the loop in Step 5 is exited when T exceeds $T_{\max}$.

What remains to be discussed are the choices of the hyperparameters $\tilde{v}$, $L_{\max}$, and $T_{\max}$. Throughout this paper, we have set $\tilde{v} = 0.75$, $L_{\max} = 10p$, and $T_{\max} = \lceil n/2 \rceil$, where $\lceil n/2 \rceil$ denotes the smallest integer that is equal to or larger than $n/2$. An explanation and a discussion of these choices are deferred to Appendix J in the supplementary materials.

4. Numerical simulations

In this section, the performances of the proposed *T-Rex* selector and the benchmark methods are compared in a simulation study. The benchmark methods in low-dimensional settings (i.e., $p \leq n$) are the well-known Benjamini–Hochberg (*BH*) method [17], the Benjamini-Yekutieli (*BY*) method [18], and the *fixed-X* knockoff methods [19], while the *model-X* knockoff methods [20] are the benchmarks in high-dimensional settings (i.e., $p > n$). Knockoff methods come in two variations called "knockoff" and "knockoff+". Only the "knockoff+" version is an FDR controlling method. For a detailed explanation and discussion of the benchmark methods, the reader is referred to Appendix H in the supplementary materials.

4.1. Setup and results

We generate a sparse high-dimensional setting[12] with n observations, p predictors, and a response given by the linear model in (1). Further, $\beta_j = 1$ for p_1 randomly selected j's while $\beta_j = 0$ for the others. The predictors are (i) sampled independently from the standard normal distribution (Figs. 5 and 6) and (ii) sampled from an autoregressive model of order one with autocorrelation coefficient $\rho = 0.5$ (Fig. 7). The standard deviation of the noise σ is chosen such that the signal-to-noise ratio (SNR), which is given by $\mathrm{Var}[\boldsymbol{X}\boldsymbol{\beta}] / \sigma^2$, is equal to the desired value. In Appendices K and L of the supplementary materials, we show results for non-Gaussian predictors and heavy-tailed noise settings. The specific values of the above described simulation setting and the parameters of the *T-Rex* selector, i.e., the values of n, p, p_1, SNR, K, L, T, v, are specified in the figure captions. The results are averaged over $MC = 955$ Monte Carlo replications.[13]

First, in order to assess the FDR control performance and the achieved power of the *T-Rex* selector, respectively, the average FDP, $\widehat{\mathrm{FDP}}$, and TPP are computed over a two-dimensional grid for v and T for different values of L. We evaluate the performance of the *T-Rex* selector in combination with the proposed extended calibration algorithm in Algorithm 2 across different sparsity levels and SNR values at a targeted FDR of 10%. That is, all other parameters remain constant while we vary the number of true active variables p_1 (i.e., different sparsity levels) and the SNR. This approach allows us to compare the performance of the *T-Rex* selector against benchmark methods in various scenarios.

The reported average FDP, $\widehat{\mathrm{FDP}}$, and TPP (all averaged over 955 Monte Carlo replications) in Figs. 5, 6, and 7 are estimates of the FDR, $\widehat{\mathrm{FDR}}$, and TPR, respectively. For this reason, the results are discussed in terms of the FDR, $\widehat{\mathrm{FDR}}$, and TPR in the captions of the figures, while the axes labels emphasize that the average FDP, $\widehat{\mathrm{FDP}}$, and TPP are plotted.

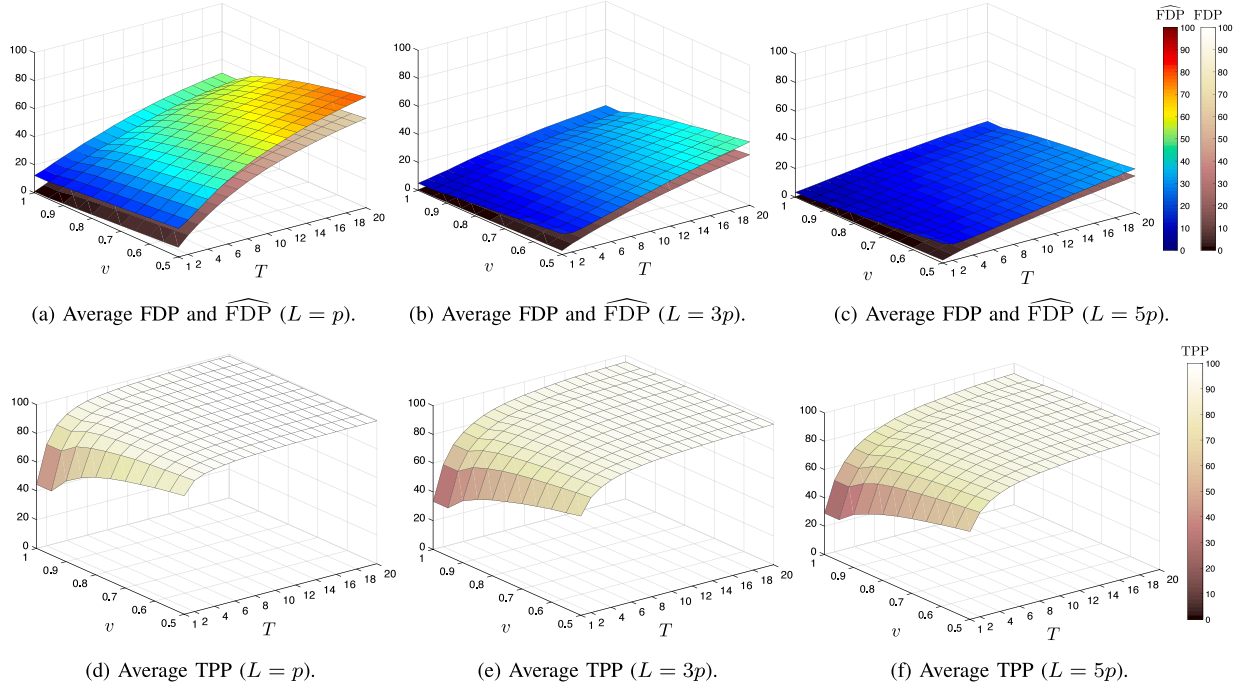
The simulation results confirm that the proposed *T-Rex* selector possesses the FDR control property. Moreover, the simulation results show that the *T-Rex* selector outperforms the benchmark methods and that its computation time is multiple orders of magnitude lower than that of its competitors (see Fig. 1 in Section 1 and Table 1). The detailed descriptions and discussions of the simulation results are given in the captions of Figs. 5, 6, and 7. Appendix K in the supplementary materials contains additional simulations that confirm the superior performance of the *T-Rex* selector under various non-Gaussian, heavy-tailed, and skewed data distributions. Furthermore, Appendix L in the supplementary materials discusses in more detail the robustness of the T-Rex selector in the presence of non-Gaussian heavy-tailed noise.

5. Simulated genome-wide association study

The *T-Rex* selector and the benchmark methods are applied to conduct a high-dimensional simulated case-control GWAS. The size of the GWAS was chosen, such that it was still practically feasible to compute the computationally intensive benchmark methods. The goal is to detect the single nucleotide polymorphisms (SNPs) that are associated with a disease of interest (i.e., active variables). At the same time, it is important to keep the number of selected SNPs that are not associated with that disease (i.e., null variables) low.

[12] Additional simulation results that allow for a performance comparison of the proposed *T-Rex* selector to the *BH* method, the *BY* method, and the *fixed-X* knockoff methods in a low-dimensional setting are deferred to Appendix I in the supplementary materials.

[13] The reason for running 955 Monte Carlo replications is that the simulations were conducted on the Lichtenberg High-Performance Computer of the Technische Universität Darmstadt, which consists of multiple nodes of 96 CPUs each. In order to run computationally efficient simulations, our computation jobs are designed to request 2 nodes and run 5 cycles on each CPU while one CPU acts as the master, i.e., $(2 \cdot 96 - 1) \cdot 5 = 955$.



(a) Average FDP and $\widehat{\text{FDP}}$ ($L = p$). (b) Average FDP and $\widehat{\text{FDP}}$ ($L = 3p$). (c) Average FDP and $\widehat{\text{FDP}}$ ($L = 5p$).

(d) Average TPP ($L = p$). (e) Average TPP ($L = 3p$). (f) Average TPP ($L = 5p$).

Fig. 5. The *T-Rex* selector controls the FDR for all values of v and T while achieving a high power, even at low values of T. Note that the FDR control is tighter for large values of L. This observation led to the development of Algorithm 2. Moreover, we observe that the conditions in Theorem 3 hold on average (i.e., ceteris paribus, $\widehat{\text{FDP}}(v, T, L)$ is monotonically decreasing in v and monotonically increasing in T). Setup: $n = 300$, $p = 1000$, $p_1 = 10$, $K = 20$, SNR $= 1$, $MC = 955$.

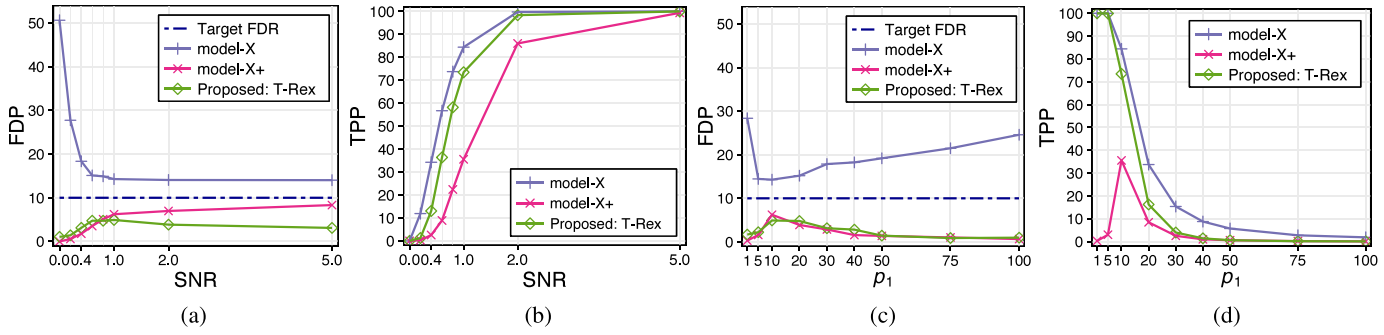


Fig. 6. General: The *model-X* knockoff method fails to control the FDR. Among the FDR-controlling methods, the *T-Rex* selector outperforms the *model-X* knockoff+ method in terms of power. **Details**: **(a)** The *T-Rex* selector and the *model-X* knockoff+ method control the FDR at a target level of 10% for the whole range of SNR values while the *model-X* knockoff method fails to control the FDR and performs poorly at low SNR values. Setup: $n = 300$, $p = 1000$, $p_1 = 10$, $T_{\max} = \lceil n/2 \rceil$, $L_{\max} = 10p$, $K = 20$, $MC = 955$. **(b)** As expected, the TPR (i.e., power) increases with respect to the SNR. It is remarkable that even though the FDR of the *T-Rex* selector lies below that of the *model-X* knockoff+ method for SNR values larger than 0.6, its power exceeds that of its strongest FDR-controlling competitor. The high power of the *model-X* knockoff method cannot be interpreted as an advantage, because the method does not control the FDR. Setup: Same as in Figure (a). **(c)** As in Figure (a), only the *T-Rex* selector and the *model-X* knockoff+ method control the FDR at a target level of 10%, whereas the *model-X* knockoff method always exceeds the target level. Setup: $n = 300$, $p = 1000$, $T_{\max} = \lceil n/2 \rceil$, $L_{\max} = 10p$, $K = 20$, SNR $= 1$, $MC = 955$. **(d)** Among the FDR-controlling methods, the *T-Rex* selector has by far the highest power for sparse settings. The power of the *model-X* knockoff method exceeds that of the FDR-controlling methods, but this cannot be interpreted as an advantage of the method since it exceeds the target FDR level. Note that for an increasing number of active variables the power drops for all methods since apparently the number of data points $n = 300$ does not suffice in the simulated settings with a low sparsity level, i.e., settings with many active variables. Setup: Same as in Figure (c).

5.1. Setup

The genotypes of 700 cases and 300 controls are simulated based on haplotypes from phase 3 of the International HapMap project [55] using the software HAPGEN2 [56]. We simulated 10 randomly selected disease loci on the first 20,000 SNPs of chromosome 15 (contains 42,351 SNPs in total) with randomly selected risk alleles (either 0 or 1 with $\mathbb{P}(\text{“0”}) = \mathbb{P}(\text{“1”}) = 0.5$) and with the heterozygote risks and the homozygote risks being sampled from the uniform distribution on the intervals $[1.5, 2]$ and $[2.5, 3]$, respectively. Since we are conducting a case-control study, the control and case phenotypes are 0 and 1, respectively. Note that the SNPs and the phenotype represent the candidate variables and the response, respectively, while the disease loci represent the indices of the active variables. Thus, we have $p_1 = 10$ active variables and $p_0 = 19{,}990$ null variables. The number of observations is $n = 1000$ (700 cases and 300 controls). The results are averaged over 100 data sets satisfying the above specifications. The detailed description of the setup and the preprocessing of the data is deferred to Appendix J in the supplementary materials.

5.2. Results

In this HAPGEN2-based GWAS benchmarking, the *T-Rex* selector demonstrates its real-life applicability, as it is the only FDR-controlling method with a positive TPR, and its sequential computation time is 4 min (vs. more than 12 h for the knockoff methods). The results

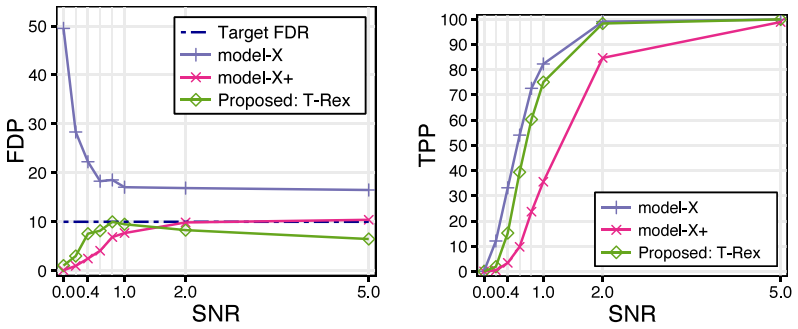


Fig. 7. Average FDP and TPP in the case of dependent predictors: The *T-Rex* selector controls the FDR, has the highest power among the FDR-controlling methods, and reaches the almost highest possible TPR level at an SNR of 2 while the *model-X* knockoff+ method requires an SNR of 5 to reach the same TPR level. The *model-X* knockoff+ method also controls the FDR except for an SNR of 5, where it slightly exceeds the target FDR, and the *model-X* knockoff method does not control the FDR. The predictors were sampled from an autoregressive model of order one (AR(1)) with Gaussian noise and an autocorrelation coefficient $\rho = 0.5$. Setup: $n = 300$, $p = 1000$, $p_1 = 10$, $T_{\max} = \lceil n/2 \rceil$, $L_{\max} = 10p$, $K = 20$, $MC = 955$.

Table 1
The proposed *T-Rex* selector is the only method whose average FDP lies below the target FDR level of 10% while achieving a non-zero power. The only competitor that provably possesses the FDR control property, namely the *model-X* knockoff+ method, has an average FDP of 0% but also an average TPP of 0%, i.e., it has no power. The *model-X* knockoff method exceeds the target FDR level. The computationally cheap procedure of plugging the marginal *p*-values into the *BH* method or the *BY* method, which has been a standard procedure in GWAS, fails in this high-dimensional setting. The sequential computation time of the proposed *T-Rex* selector in combination with the extended calibration algorithm in Algorithm 2 is roughly 4 min as compared to more than 12.5 h for the *model-X* methods. That is, the *T-Rex* selector is 183 times faster than its strongest competitors. Note that this is only a comparison of the sequential computation times. Since the random experiments of the proposed *T-Rex* selector are independent and, therefore, can be run in parallel on multicore computers, an additional substantial speedup can be achieved.

Methods	FDR control?	Average FDP (in %)	Average TPP (in %)	Average sequential computation time (hh:mm:ss)	Average relative sequential computation time
Proposed: *T-Rex*	✓	**6.45**	**38.50**	**00:04:05**	**1**
model-X+	✓	0.00	0.00	12:32:47	183.71
model-X	✗	13.07	41.40	12:32:47	183.71
BY	✗	94.00	0.00	00:00:00	0.00
BH	✗	99.00	0.00	00:00:00	0.00

(i.e., FDR, TPR, and sequential computation time) and a discussion thereof are given in Table 1, while additional results are deferred to Appendix J in the supplementary materials.

6. Conclusion

The *T-Rex* selector, a new fast FDR-controlling variable selection framework for high-dimensional settings, was proposed and benchmarked against existing methods in numerical simulations and a simulated GWAS. The *T-Rex* selector is, to the best of our knowledge, the first multivariate high-dimensional FDR-controlling method that scales to millions of variables in a reasonable amount of computation time. Since the *T-Rex* random experiments can be computed in parallel, multicore computers allow for additional substantial savings in computation time. These properties make the *T-Rex* selector a suitable method especially for large-scale GWAS.

In order to ensure the reproducibility of the presented results and enhance the usability of the proposed *T-Rex* selector, the actively maintained open-source R software package TRexSelector has been made available on CRAN [49].

A current limitation of the *T-Rex* selector concerns its high random-access memory (RAM) usage for storing dummies. This issue has been alleviated through the use of sophisticated memory mapping technologies, enabling efficient use of the solid-state drive (SSD) of a computer to virtually extend the available RAM on standard laptops, as detailed in [43]. Nevertheless, advancements to reduce memory demand and enhance computational efficiency are a priority in our future research to allow for an even further scalability of the framework to potentially billions of variables and, thus, allow for FDR-controlled multivariate association testing using whole genome sequencing data.

Moreover, the developed R package TRexSelector currently does not allow for complex valued input data. We are working on extending the software package to handle complex valued data, thereby expanding the applicability of the *T-Rex* selector to a broader range of applications in various scientific and engineering domains.

As a next step, we shall conduct multiple reproducibility studies applying the *T-Rex* selector on large-scale genotype and phenotype data from the UK Biobank [57] in order to reproduce some of the reported results in the GWAS catalog [12]. Our aim is to confirm past discoveries, discover new genetic associations, and flag potentially false reported genetic associations. We plan to publish our results as a curated catalog of reproducible genetic associations and hope that this endeavor helps scientists to focus their efforts in revealing the causal mechanisms behind the genetic associations on the most promising and reproducible genetic associations.

CRediT authorship contribution statement

Jasin Machkour: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michael Muma:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Daniel P. Palomar:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Michael Fauss, Taulant Koka, and Fabian Scheidt for many discussions and helpful feedback. We also thank Simon Tien for his help in developing the R software packages *TRexSelector* and *tlars*. Extensive computations on the Lichtenberg High-Performance Computer of the Technische Universität Darmstadt were conducted for this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.sigpro.2025.109894.

Data availability

The open source R package TRexSelector containing the implementation of the T-Rex selector is available on CRAN.

References

[1] P.-J. Chung, J.F. Bohme, C.F. Mecklenbrauker, A.O. Hero, Detection of the number of signals using the Benjamini-Hochberg procedure, IEEE Trans. Signal Process. 55 (6) (2007) 2497–2508.
[2] J. Chen, W. Zhang, H.V. Poor, A false discovery rate oriented approach to parallel sequential change detection problems, IEEE Trans. Signal Process. 68 (2020) 1823–1836.
[3] Z. Chen, F. Sohrabi, W. Yu, Sparse activity detection for massive connectivity, IEEE Trans. Signal Process. 66 (7) (2018) 1890–1904.
[4] Z. Tan, Y.C. Eldar, A. Nehorai, Direction of arrival estimation using co-prime arrays: A super resolution viewpoint, IEEE Trans. Signal Process. 62 (21) (2014) 5565–5576.
[5] P. Di Lorenzo, A.H. Sayed, Sparse distributed learning based on diffusion adaptation, IEEE Trans. Signal Process. 61 (6) (2012) 1419–1433.
[6] K. Benidis, Y. Feng, D.P. Palomar, Sparse portfolios for high-dimensional financial index tracking, IEEE Trans. Signal Process. 66 (1) (2017) 155–170.
[7] A.M. Zoubir, V. Koivunen, Y. Chakhchoukh, M. Muma, Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts, IEEE Signal Process. Mag. 29 (4) (2012) 61–80.
[8] A.M. Zoubir, V. Koivunen, E. Ollila, M. Muma, Robust Statistics for Signal Processing, Cambridge Univ. Press, 2018.
[9] J. Machkour, B. Alt, M. Muma, A.M. Zoubir, The outlier-corrected-data-adaptive Lasso: A new robust estimator for the independent contamination model, in: Proc. 25th Eur. Signal Process. Conf., EUSIPCO, 2017, pp. 1649–1653.
[10] J. Machkour, M. Muma, B. Alt, A.M. Zoubir, A robust adaptive Lasso estimator for the independent contamination model, Signal Process. 174 (2020) 107608.
[11] C. Yang, X. Shen, H. Ma, B. Chen, Y. Gu, H.C. So, Weakly convex regularized robust sparse recovery methods with theoretical guarantees, IEEE Trans. Signal Process. 67 (19) (2019) 5046–5061.
[12] A. Buniello, J.A.L. MacArthur, M. Cerezo, L.W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, et al., The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, Nucleic Acids Res. 47 (D1) (2019) D1005–D1012.
[13] S.J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D.J. Hunter, G. Thomas, J.N. Hirschhorn, G. Abecasis, D. Altshuler, J.E. Bailey-Wilson, et al., Replicating genotype–phenotype associations, Nature 447 (7145) (2007) 655–660.
[14] P.M. Visscher, N.R. Wray, Q. Zhang, P. Sklar, M.I. McCarthy, M.A. Brown, J. Yang, 10 years of GWAS discovery: Biology, function, and translation, Am. J. Hum. Genet. 101 (1) (2017) 5–22.
[15] J.E. Huffman, Examining the current standards for genetic discovery and replication in the era of mega-biobanks, Nature Commun. 9 (1) (2018) 1–4.
[16] M.D. Gallagher, A.S. Chen-Plotkin, The post-GWAS era: From association to function, Am. J. Hum. Genet. 102 (5) (2018) 717–730.
[17] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B. Stat. Methodol. 57 (1) (1995) 289–300.
[18] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Ann. Statist. 29 (4) (2001) 1165–1188.
[19] R.F. Barber, E.J. Candès, Controlling the false discovery rate via knockoffs, Ann. Statist. 43 (5) (2015) 2055–2085.
[20] E.J. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection, J. R. Stat. Soc. Ser. B. Stat. Methodol. 80 (3) (2018) 551–577.
[21] Z. Ren, Y. Wei, E. Candès, Derandomizing knockoffs, J. Amer. Statist. Assoc. (2021) 1–11.
[22] Z. Ren, R.F. Barber, Derandomised knockoffs: Leveraging e-values for false discovery rate control, J. R. Stat. Soc. Ser. B. Stat. Methodol. 86 (1) (2024) 122–154.
[23] N. Meinshausen, P. Bühlmann, Stability selection, J. R. Stat. Soc. Ser. B. Stat. Methodol. 72 (4) (2010) 417–473.
[24] R.D. Shah, R.J. Samworth, Variable selection with error control: Another look at stability selection, J. R. Stat. Soc. Ser. B. Stat. Methodol. 75 (1) (2013) 55–80.
[25] D.R. Cox, A note on data-splitting for the evaluation of significance levels, Biometrika 62 (2) (1975) 441–444.
[26] L. Wasserman, K. Roeder, High dimensional variable selection, Ann. Statist. 37 (5A) (2009) 2178–2201.
[27] N. Meinshausen, L. Meier, P. Bühlmann, P-values for high-dimensional regression, J. Amer. Statist. Assoc. 104 (488) (2009) 1671–1681.
[28] R.F. Barber, E.J. Candès, A knockoff filter for high-dimensional selective inference, Ann. Statist. 47 (5) (2019) 2504–2537.
[29] R. Lockhart, J. Taylor, R.J. Tibshirani, R. Tibshirani, A significance test for the Lasso, Ann. Statist. 42 (2) (2014) 413–468.
[30] W. Fithian, D. Sun, J. Taylor, Optimal inference after model selection, 2014, arXiv preprint, arXiv:1410.2597.
[31] J.D. Lee, D.L. Sun, Y. Sun, J.E. Taylor, Exact post-selection inference, with application to the Lasso, Ann. Statist. 44 (3) (2016) 907–927.
[32] R.J. Tibshirani, J. Taylor, R. Lockhart, R. Tibshirani, Exact post-selection inference for sequential regression procedures, J. Amer. Statist. Assoc. 111 (514) (2016) 600–620.
[33] A.J. Miller, Selection of subsets of regression variables, J. R. Stat. Soc. Ser. A. Gen. 147 (3) (1984) 389–410.
[34] A.J. Miller, Subset Selection in Regression, CRC Press, 2002.
[35] Y. Wu, D.D. Boos, L.A. Stefanski, Controlling variable selection by the addition of pseudovariables, J. Amer. Statist. Assoc. 102 (477) (2007) 235–243.
[36] D. Williams, Probability with Martingales, Cambridge Univ. Press, 1991.
[37] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B. Stat. Methodol. 58 (1) (1996) 267–288.
[38] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2) (2004) 407–499.
[39] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B. Stat. Methodol. 67 (2) (2005) 301–320.
[40] H. Zou, The adaptive Lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (476) (2006) 1418–1429.
[41] J. Machkour, M. Muma, D.P. Palomar, High-dimensional false discovery rate control for dependent variables, 2024, arXiv preprint arXiv:2401.15796.
[42] J. Machkour, M. Muma, D.P. Palomar, False discovery rate control for fast screening of large-scale genomics biobanks, in: Proc. 22nd IEEE Statist. Signal Process. Workshop, SSP, 2023, pp. 666–670.
[43] F. Scheidt, J. Machkour, M. Muma, Solving FDR-controlled sparse regression problems with five million variables on a laptop, in: Proc. IEEE 9th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process., CAMSAP, 2023, pp. 116–120.
[44] J. Machkour, M. Muma, D.P. Palomar, False discovery rate control for grouped variable selection in high-dimensional linear models using the T-Knock filter, in: 30th Eur. Signal Process. Conf., EUSIPCO, 2022, pp. 892–896.
[45] J. Machkour, M. Muma, D.P. Palomar, The informed elastic net for fast grouped variable selection and FDR control in genomics research, in: Proc. IEEE 9th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process., CAMSAP, 2023, pp. 466–470.
[46] T. Koka, J. Machkour, M. Muma, False discovery rate control for Gaussian graphical models via neighborhood screening, in: Proc. 32nd Eur. Signal Process. Conf., EUSIPCO, 2024, pp. 2482–2486.
[47] J. Machkour, A. Breloy, M. Muma, D.P. Palomar, F. Pascal, Sparse PCA with false discovery rate controlled variable selection, in: Proc. IEEE 49th Int. Conf. Acoust. Speech Signal Process, ICASSP, 2024, pp. 9716–9720.
[48] J. Machkour, D.P. Palomar, M. Muma, FDR-controlled portfolio optimization for sparse financial index tracking, 2024, arXiv preprint arXiv:2401.15139.
[49] J. Machkour, S. Tien, D.P. Palomar, M. Muma, TRexSelector: T-Rex selector: High-dimensional variable selection & FDR control, 2024, R package version 1.0.0. [Online]. Available: https://CRAN.R-project.org/package=TRexSelector.
[50] J. Machkour, S. Tien, D.P. Palomar, M. Muma, tlars: The T-LARS algorithm: Early-terminated forward variable selection, 2024, R package version 1.0.1. [Online]. Available: https://CRAN.R-project.org/package=tlars.
[51] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused Lasso, J. R. Stat. Soc. Ser. B. Stat. Methodol. 67 (1) (2005) 91–108.
[52] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, Ann. Appl. Stat. 1 (2) (2007) 302–332.
[53] J.D. Storey, The positive false discovery rate: A Bayesian interpretation and the q-value, Ann. Statist. 31 (6) (2003) 2013–2035.
[54] W. Su, M. Bogdan, E.J. Candès, False discoveries occur early on the Lasso path, Ann. Statist. 45 (5) (2017) 2133–2150.
[55] T.I.H.. Consortium, Integrating common and rare genetic variation in diverse human populations, Nature 467 (7311) (2010) 52–58.
[56] Z. Su, J. Marchini, P. Donnelly, HAPGEN2: Simulation of multiple disease SNPs, Bioinformatics 27 (16) (2011) 2304–2305.
[57] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al., UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age, PLoS Med. 12 (3) (2015) e1001779.