A BOOTSTRAP MODEL COMPARISON TEST FOR IDENTIFYING GENES WITH CONTEXT-SPECIFIC PATTERNS OF GENETIC REGULATION

BY MYKHAYLO M. MALAKHOV[1,a], BEN DAI[2,c], XIAOTONG T. SHEN[3,d], AND WEI PAN[1,b]

[1]*Division of Biostatistics, University of Minnesota,* [a]*malak039@umn.edu;* [b]*panxx014@umn.edu*

[2]*Department of Statistics, The Chinese University of Hong Kong,* [c]*bendai@cuhk.edu.hk*

[3]*School of Statistics, University of Minnesota,* [d]*xshen@umn.edu*

Understanding how genetic variation affects gene expression is essential for a complete picture of the functional pathways that give rise to complex traits. Although numerous studies have established that many genes are differentially expressed in distinct human tissues and cell types, no tools exist for identifying the genes whose expression is differentially regulated. Here we introduce DRAB (Differential Regulation Analysis by Bootstrapping), a gene-based method for testing whether patterns of genetic regulation are significantly different between tissues or other biological contexts. DRAB first leverages the elastic net to learn context-specific models of local genetic regulation and then applies a novel bootstrap-based model comparison test to check their equivalency. Unlike previous model comparison tests, our proposed approach can determine whether population-level models have equal predictive performance by accounting for the variability of feature selection and model training. We validated DRAB on mRNA expression data from a variety of human tissues in the Genotype-Tissue Expression (GTEx) Project. DRAB yielded biologically reasonable results and had sufficient power to detect genes with tissue-specific regulatory profiles while effectively controlling false positives. By providing a framework that facilitates the prioritization of differentially regulated genes, our study enables future discoveries on the genetic architecture of molecular phenotypes.

1. Introduction. Differential gene expression analysis is among the most common techniques in computational biology. It is used to compare the RNA transcript abundance of a gene between two different biological contexts or experimental conditions, and to determine whether the difference is greater than what would be expected due to natural random variation. A wide variety of statistical tools have been developed for differential expression analysis, from simple t-tests to sophisticated Bayesian approaches (Pan (2002); Bullard et al. (2010); Soneson and Delorenzi (2013); Seyednasrollah, Laiho and Elo (2015)). When applied to tissue-specific gene expression measurements, such as those available from the Genotype-Tissue Expression (GTEx) Project (Lonsdale et al. (2013); The GTEx Consortium (2020)), differential expression analysis can identify a set of genes whose expression is significantly different between two given tissues.

Gene expression studies can also provide insight into the genetic architecture of complex human diseases. Indeed, many recent approaches have sought to uncover causal loci by considering the functional pathways through which genes affect traits (Gallagher and Chen-Plotkin (2018); Wong et al. (2021)). A notable example is the transcriptome-wide association study (TWAS) framework, which prioritizes disease-relevant genes by testing for association with genetically regulated expression levels (Gamazon et al. (2015); Gusev et al. (2016)).

Keywords and phrases: differential regulation, gene expression, model selection test, penalized regression, transcriptome prediction.

TWAS has been applied in over a thousand studies and has led to a number of innovative methodological extensions (Barbeira et al. (2018); Hu et al. (2019); Yang et al. (2020); Lin et al. (2022)). The GTEx Project, for instance, utilized TWAS and related methods to characterize genetic effects on the transcriptome across human tissues and to link these regulatory mechanisms to trait and disease associations (The GTEx Consortium (2020); Barbeira et al. (2021)).

Yet despite the ubiquity of differential expression analysis and the importance of understanding how genetic regulation induces differences in gene expression, a fundamental problem remains poorly studied: when does the genetic regulation of expression differ between various tissue and cellular contexts?

This question is challenging to answer for at least two reasons. First, differential regulation is distinct from differential expression. For some genes, differences in expression levels are driven by fluctuations in tissue or cell microenvironment rather than by differences in networks of genetic regulatory activity. As a result, the presence of differential gene expression does not necessarily imply that context-specific patterns of genetic regulation exist. By the same reasoning, differentially regulated genes may not actually be differentially expressed. Second, the genetic architecture of gene expression is complex. Whereas context-specific gene expression levels can be easily quantified, genetic regulatory profiles might differ between biological contexts in a myriad ways. For example, consider two human tissues: different variants might be responsible for regulating gene expression in each tissue, or the same variants might have different effect sizes, or one tissue might be enriched in expression quantitative trait loci (eQTLs) relative to the other. Therefore, patterns of genetic regulation cannot be summarized by a single index or compared using classical statistical tests.

In this paper we introduce DRAB (Differential Regulation Analysis by Bootstrapping), a statistical method for determining whether gene-level genetic regulation is significantly different between two tissues or other biological contexts. The premise of our approach is to train penalized regression models that capture the context-specific effects of local genetic variation on gene expression, and then to statistically test whether those models have equal predictive performance on an independent test set. During the training process each model learns the particular regulatory patterns present in the context that it is trained on, and so testing for model equivalency effectively tests whether the genetic regulation of expression is equivalent between those contexts. To perform such a test without conditioning on a specific pair of trained models, we develop a novel bootstrap-based model comparison framework. The DRAB test accounts for the uncertainty of performing feature selection and model fitting on randomly sampled training sets, allowing us to interpret it as a test of equivalency between population-level models of genetic regulation. As such, DRAB directly addresses both of the challenges outlined above. To our knowledge this is the first statistical method for identifying genes with tissue-specific (or other context-specific) patterns of genetic regulation.

We demonstrate DRAB by applying it to whole-genome sequencing and mRNA gene expression data from 23 selected human tissues in the GTEx v8 data release (The GTEx Consortium (2020)). We show that our method is able to control type I error and has sufficient power to detect tissue-specific differences in genetic regulatory profiles, while conventional model comparison tests yield many spurious results. By evaluating a broad array of tissue pairs, we confirm that the genetic regulation of gene expression is more similar in biologically related tissues (e.g. different brain tissues) than in unrelated tissue pairs (e.g. blood and brain tissues). In particular, we identify sets of genes whose expression is differentially regulated between blood and various brain regions, between distinct nervous system tissues, and between a variety of other selected tissue pairs. To facilitate further discovery we make an R implementation of DRAB available on GitHub, along with supporting bash scripts for data processing (`https://github.com/MykMal/drab`).

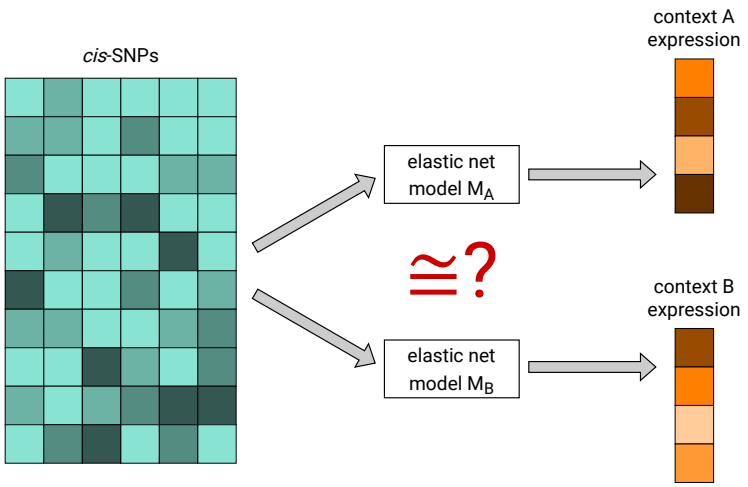

FIG 1. *The DRAB method is comprised of two stages. First, elastic net models are trained to predict gene expression from cis-SNPs in each context. Second, the context-specific models are tested for statistical equivalency at the population level using a bootstrap-based model comparison test on an independent test set.*

2. Methods.

2.1. *Overview of DRAB.* Consider a gene g, and let A and B denote two tissues, disease states, or other biological sample types. The purpose of DRAB is to test the null hypothesis that the *cis*-genetic regulation of g is equivalent in context A and context B. To do this, DRAB first trains elastic net models in each context to predict the expression of g from single nucleotide polymorphisms (SNPs) located near the gene. The training algorithm simultaneously performs feature selection and model fitting, so that the resulting models capture context-specific patterns of genetic regulation in *cis* for gene g. Then DRAB tests whether those context-specific models are statistically equivalent using a novel model comparison test on an independent test set. Our proposed model comparison test differs from previous approaches in that it accounts for the uncertainty of learning a regression function on randomly sampled training sets, which we estimate through bootstrap resampling. The two stages of DRAB are visualized in Figure 1 and detailed below. In practice, DRAB automatically repeats this process for each gene on a user-specified list of genes. Note that individual-level genotype and gene expression data are required.

2.2. *Stage 0: data pre-processing.* Let D_A be the training set for context A, D_B be the training set for context B, and D_T be the test set. In subsequent stages of DRAB, D_A and D_B are used for training context-specific models of genetic regulation, while D_T is used for comparing the two models' predictive performance. To ensure a fair comparison, we assume that $|D_A| = |D_B|$ and $D_T \cap (D_A \cup D_B) = \emptyset$. That is, the two training sets must have the same cardinality and each be independent from the test set. The training and test sets do not need to be pre-specified when running DRAB, however, since the program will automatically create the largest possible training sets D_A, D_B from the provided input data and then set aside the remaining data for the test set D_T. The only requirement in practice is that the input data for contexts A and B are not completely overlapping, since full overlap would make it impossible to form an independent test set.

DRAB begins by extracting SNPs that are located within the *cis*-locus of g, which we define as ranging from 500 kilobase pairs upstream of the gene's transcription start site to

500 kilobase pairs downstream of its transcription end site. We refer to the variants within this region as *cis*-SNPs. Then multiple linear regression is performed to adjust the expression of g and the genotypes of each *cis*-SNP for a set of user-specified covariates. These regression models are fit separately on the sets D_A, D_B, and D_T to preserve their independence. Thus, a total of $3(p+1)$ models are fit, where p is the number of *cis*-SNPs belonging to g. Next we center and standardize the expression residuals, as well as the residuals of each *cis*-SNP, to have a mean of zero and a variance of one. As a last step, we remove any monomorphic *cis*-SNPs. The normalized residuals are subsequently used as the expression and genotype values in the main stages of DRAB.

2.3. *Stage 1: model training.* DRAB independently trains two parametric machine learning models: model M_A and model M_B. Model M_A is trained to predict the (adjusted) expression levels of gene g in context A, using data from the individuals in training set D_A. Similarly, model M_B is trained to predict the (adjusted) expression levels of gene g in context B, using data from the individuals in training set D_B. Both models are given the same set of (adjusted) *cis*-SNPs as features, but differ in their response variables and training sets.

In particular, we use Gaussian family elastic net models (Zou and Hastie (2005)). The elastic net is a penalized regression method that linearly combines the L_1 norm from lasso regression (Tibshirani (1996)) and the L_2 norm from ridge regression (Hoerl and Kennard (1970)). Consequently, the elastic net performs simultaneous regularization and variable selection, making it appropriate for learning sparse models from high-dimensional data with correlated features. Given genotype dosage vectors $x_i \in \mathbb{R}^p$ and gene expression measurements $y_i \in \mathbb{R}$ for individuals $i = 1, \ldots, N$, the elastic net minimizes the objective function

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - x^T\beta\right)^2 + \lambda\left(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right), \tag{1}$$

where $\lambda \geq 0$ is a complexity parameter and $0 \leq \alpha \leq 1$ controls the relative weight of each penalty. We fix the mixing parameter α at $\alpha = 0.5$, which tends to either select or leave out entire groups of correlated features (Zou and Hastie (2005)). As a result, M_A and M_B each select the groups of *cis*-eQTLs that are specific to the contexts they are trained on. By performing both feature selection and model fitting, the elastic net efficiently learns tissue-specific patterns of *cis*-genetic regulation.

To train the elastic net models, DRAB relies on the R package `glmnet` (Friedman, Hastie and Tibshirani (2010); Tibshirani et al. (2012)). The `glmnet` package dynamically computes a grid of values for the tuning parameter λ, ensuring that the grid covers the entire range of possible solutions. Then we use 5-fold cross validation to select the value of λ that yields the most parsimonious model whose cross-validated error is within one standard deviation of the minimum. DRAB utilizes this approach instead of selecting the best possible λ because sparse models tend to outperform more polygenic models at transcriptome prediction (Wheeler et al. (2016); Fryett, Morris and Cordell (2020)).

Note that other algorithms for regression and feature selection could be justifiably used here instead of a linear model with the elastic net penalty. Our proposed model comparison test relies on empirical distributions and can thus be performed with any parametric regression models. However, the elastic net algorithm offers several desirable properties that make it the most natural choice for our application. Most importantly, the elastic net selects sparse groups of context-specific SNPs and learns context-specific weights for each selected SNP. This makes DRAB robust to the high dimensionality of genetic architectures and the inherent complexity of linkage disequilibrium between eQTLs involved in transcriptional regulation. Moreover, the elastic net is commonly used for transcriptome imputation in TWAS (Gamazon et al. (2015); Barbeira et al. (2018); Gillies et al. (2018); Wu et al. (2018); Guo et al.

(2021)) and previous research has shown that it outperforms several other machine learning algorithms at this task (Okoro et al. (2021)).

2.4. *Stage 2: testing model equivalency.*

2.4.1. *Limitations of conditional approaches.* Our goal is to conduct a model comparison test in order to determine whether the stage 1 models M_A and M_B are statistically equivalent at the population level with respect to some loss function L. The standard practice for model comparison is to evaluate both models on an independent test set and compare their prediction errors. When a formal decision is desired, one can test the null hypothesis that the prediction errors are equal. For regression learning tasks, in particular, a paired samples t-test on the squared residual errors is used. Here we explain why such approaches are inadequate for assessing whether context-specific models of genetic regulation are equivalent.

To distinguish the fitted stage 1 models from the theoretical population-level models, we will use the notation $\widehat{M}_j = \widehat{M}_j(D_j)$ to signify the model trained on set D_j for $j = A, B$. We assume that the trained models are consistent for the true models: $\widehat{M}_j \xrightarrow{p} M_j$ as $|D_j| \to \infty$ for $j = A, B$. That is, we assume that the finite-sample fitted models converge in probability to the true population-level models as the training sample sizes increase to infinity. Now let $L_i(\widehat{M}_j; D_T)$ denote the squared prediction residual for model $\widehat{M}_j$ and individual $i \in D_T$. Then the average difference in prediction errors between model $\widehat{M}_A$ and model $\widehat{M}_B$ on the test set D_T is

$$\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T) = \frac{\sum_{i=1}^{|D_T|} \left(L_i(\widehat{M}_A; D_T) - L_i(\widehat{M}_B; D_T) \right)}{|D_T|}. \tag{2}$$

The paired samples t-test relies on three assumptions: prediction errors are "paired" between the two models, each pair is independent of the rest, and the pairwise differences are approximately normally distributed. The first assumption is satisfied because prediction errors for both models are computed on the same test set, so $L_i(\widehat{M}_A; D_T)$ and $L_i(\widehat{M}_B; D_T)$ form a pair for each individual $i \in D_T$. We follow the precedent in large-scale genetics studies of assuming that genotyped individuals are unrelated, which ensures that the second assumption is satisfied. Finally, the third assumption follows from the central limit theorem for a sufficiently large test set.

The paired samples t-test statistic is given by

$$T = \frac{\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)}{\sqrt{\widehat{Var}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T) \right)}}. \tag{3}$$

Variance can be estimated in the standard way as

$$\widehat{Var}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T) \right) = \frac{\sum_{i=1}^{|D_T|} \left(\widehat{\Delta L}_i - \overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T) \right)^2}{|D_T|(|D_T| - 1)} \tag{4}$$

where $\widehat{\Delta L}_i = L_i(\widehat{M}_A; D_T) - L_i(\widehat{M}_B; D_T)$. Under the null hypothesis that models $\widehat{M}_A$ and $\widehat{M}_B$ have equal predictive performance with respect to L, the statistic T follows a Student's t-distribution with $|D_T| - 1$ degrees of freedom.

Although this model comparison framework is commonly used in the statistical learning literature, it has a major limitation. By definition, the t-test described above determines whether the mean difference between prediction errors for model $\widehat{M}_A$ and model $\widehat{M}_B$ is zero.

Importantly, it pertains to the fitted models $\widehat{M}_A, \widehat{M}_B$ rather than the theoretical population-level models M_A, M_B. For some applications, conditioning on the trained models is satisfactory or even desired; but in our case ignoring the uncertainty associated with model training would lead to a greatly inflated false positive rate.

It is well known that models pick up finite-sample errors during the training process. For example, fitting the same linear regression model on two separate data sets will yield different predictor coefficients—even if both data sets are randomly sampled from a common population. The stochastic nature of data sampling becomes especially problematic when performing feature selection, since the same feature selection algorithm might choose different sets of features when applied to different random samples from the same population (Kalousis, Prados and Hilario (2007); Nogueira, Sechidis and Brown (2018)). The issue of model training uncertainty is further exacerbated in algorithms that have a stochastic component, such as cross validation. Such algorithms are known to select different sets of features when repeatedly trained on exactly the same data set (Cadiou and Slama (2021)). Therefore, the elastic net algorithm that we use in stage 1 suffers from instability both due to the uncertainty of tuning the λ hyper-parameter using cross validation and the uncertainty of estimating coefficients on randomly sampled finite training sets.

As a result of these dual uncertainties associated with performing feature selection and model fitting, the trained stage 1 models $\widehat{M}_A$ and $\widehat{M}_B$ can be significantly different even when the theoretical models M_A and M_B are equivalent. The aim of DRAB is to compare the population-level patterns of genetic regulation between contexts A and B, so testing for equivalency between the trained models $\widehat{M}_A$ and $\widehat{M}_B$ would lead to erroneous conclusions. In fact, we show in Section 3.2 that attempting to use the standard t-test for comparing models of genetic regulation yields unacceptable results.

2.4.2. *The DRAB model comparison test.* To mitigate the problem of model training variability, we developed a new model comparison test based on the non-parametric bootstrap. Unlike the standard model comparison approach described above, our test explicitly accounts for the uncertainty of feature selection and model fitting. Hence, it can be interpreted as testing whether the theoretical population-level models, M_A and M_B, are statistically equivalent. Our proposed test is inspired by recent work on hypothesis testing for black-box models (Wasserman, Ramdas and Balakrishnan (2020); Dai, Shen and Pan (2022)), but it is methodologically novel and conceptually distinct.

The key insight behind the DRAB model comparison test is that we avoid conditioning on the fitted models $\widehat{M}_A$ and $\widehat{M}_B$, instead estimating what the paired samples t-test statistic would be under the true models of genetic regulation. Formally, we wish to test the null hypothesis

$$H_0 : E_{D_T}\left[\overline{\Delta L}(M_A, M_B; D_T)\right] = 0, \tag{5}$$

which states that the population means of the prediction errors of models M_A, M_B are equal with respect to the loss function L. It is intuitive to imagine the following statistic for this test:

$$T_\infty = \frac{\overline{\Delta L}(M_A, M_B; D_T)}{\sqrt{Var\left(\overline{\Delta L}(M_A, M_B; D_T)\right)}}. \tag{6}$$

But the theoretical population-level models M_A and M_B are unknown by definition! We therefore resort to approximating the numerator of T_∞ and its variance.

Since $\widehat{M}_A, \widehat{M}_B$ are consistent for M_A, M_B by our earlier assumption, the naive estimator for $\overline{\Delta L}(M_A, M_B; D_T)$ will yield good results as long as the training data sets are sufficiently

large. Thus, we approximate the numerator of T_∞ by substituting in the full-sample trained models:

$$\overline{\Delta L}(M_A, M_B; D_T) \approx \overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T). \tag{7}$$

Now let $\overline{\Delta L}^{*(1)}, \ldots, \overline{\Delta L}^{*(K)}$ be K non-parametric bootstrapped replicates of $\overline{\Delta L}$. That is, for each $k = 1, \ldots, K$ we create the sets $D_A^{*(k)}$ and $D_B^{*(k)}$ by randomly sampling with replacement from D_A and D_B, respectively, such that $|D_A^{*(k)}| = |D_A|$ and $|D_B^{*(k)}| = |D_B|$. Then on each set $D_A^{*(k)}$ we train a model to predict expression in context A, and denote it by $\widehat{M}_A^{*(k)} = \widehat{M}_A^{*(k)}(D_A^{*(k)})$. Similarly, on each set $D_B^{*(k)}$ we train a model to predict expression in context B, and denote it by $\widehat{M}_B^{*(k)} = \widehat{M}_B^{*(k)}(D_B^{*(k)})$. Finally, for each k we compute

$$\overline{\Delta L}^{*(k)} = \overline{\Delta L}\left(\widehat{M}_A^{*(k)}, \widehat{M}_B^{*(k)}; D_T\right). \tag{8}$$

The resulting empirical distribution of $\overline{\Delta L}^{*(1)}, \ldots, \overline{\Delta L}^{*(K)}$ approximates the sampling distribution of $\overline{\Delta L}$ across all possible context-specific training sets (Efron and Tibshirani (1993)).

To estimate the variance of $\overline{\Delta L}(M_A, M_B; D_T)$, we decompose it using the law of total variance. Observe that

$$\begin{aligned}
&Var_{D_T}\big(\overline{\Delta L}\,(M_A, M_B; D_T)\big) = \\
&\quad = E_{D_A,D_B}\left[Var_{D_T}\left(\overline{\Delta L}(M_A, M_B; D_T)\middle| M_A = \widehat{M}_A, M_B = \widehat{M}_B\right)\right] + \\
&\qquad + Var_{D_A,D_B}\left(E_{D_T}\left[\overline{\Delta L}(M_A, M_B; D_T)\middle| M_A = \widehat{M}_A, M_B = \widehat{M}_B\right]\right) \\
&\quad = E_{D_A,D_B}\left[Var_{D_T}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)\right)\right] + \\
&\qquad + Var_{D_A,D_B}\left(E_{D_T}\left[\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)\right]\right) \\
&\quad \approx Var_{D_T}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)\right) + Var_{D_A,D_B}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)\right) \\
&\quad \approx \widehat{Var}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\overline{\Delta L}^{*(k)} - \frac{1}{K}\sum_{j=1}^{K}\overline{\Delta L}^{*(j)}\right)^2.
\end{aligned} \tag{9}$$

Note that in the last line we used the bootstrap distribution of $\overline{\Delta L}$ to approximate the variance of $\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)$ across all possible training sets. Thus, in the denominator of T_∞ we estimate

$$\begin{aligned}
&Var\big(\overline{\Delta L}(M_A, M_B\,; D_T)\big) \approx \\
&\quad \approx \widehat{Var}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\overline{\Delta L}^{*(k)} - \frac{1}{K}\sum_{j=1}^{K}\overline{\Delta L}^{*(j)}\right)^2.
\end{aligned} \tag{10}$$

This variance estimate can be interpreted as the sum of two sources of uncertainty. The first term accounts for the uncertainty of the estimator $\overline{\Delta L}$ computed on the test set D_T, so it will converge to zero as $|D_T| \to \infty$. The second term, on the other hand, accounts for the uncertainty of the estimated (i.e. trained) models $\widehat{M}_A$ and $\widehat{M}_B$, so it will converge to zero only when $|D_A| \to \infty$ and $|D_B| \to \infty$. In other words, the first term pertains to the sampling

variability of the test set D_T while the second term pertains to the sampling variabilities of the training sets D_A and D_B. Including both sources of variance is the primary innovation of the DRAB test, which allows it to overcome the limitations of model comparison tests that condition on a specific pair of fitted models.

Putting everything together, our proposed DRAB test statistic is

$$T_{DRAB} = \frac{\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)}{\widehat{Var}\left(\overline{\Delta L}(\widehat{M}_A, \widehat{M}_B; D_T)\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\overline{\Delta L}^{*(k)} - \frac{1}{K}\sum_{j=1}^{K}\overline{\Delta L}^{*(j)}\right)^2}. \tag{11}$$

We claim that under the null hypothesis stated in Equation 5, the T_{DRAB} statistic approximately follows a Student's t-distribution with $|D_T| - 1$ degrees of freedom. Our proposed statistical test expressly accounts for the variability in estimating M_A and M_B through model training, and so it can be used to determine whether the true theoretical models are statistically equivalent. If the bootstrap-based DRAB test rejects the null hypothesis of model equivalency, we conclude that the patterns of genetic regulation for gene g are significantly different in contexts A and B.

3. Results.

3.1. *Data preparation.* To validate DRAB and demonstrate its use, we applied it on data from the GTEx v8 release (The GTEx Consortium (2020)). In particular, we downloaded individual-level whole-genome sequencing data as a Variant Call Format (VCF) file from the database of Genotypes and Phenotypes (dbGaP), under accession number phs000424.v8.p2. Extensive sample quality control procedures, variant quality control procedures, and read-aware phasing had already been performed on the genotype data as described in The GTEx Consortium (2020), leaving 838 individuals and 46,526,292 genomic sites. We also downloaded tissue-specific matrices of bulk gene expression from the GTEx Portal (`https://gtexportal.org/home/datasets`). The gene expression data had already been quantified, processed, and normalized as described in The GTEx Consortium (2020) and was reported in terms of Transcripts Per Million (TPM) units. Data for 49 human tissues was available, with the number of quantified genes ranging from 20,249 in whole blood to 35,054 in testis.

Prior to applying our method, we performed some additional processing on the genotype data. We first subset the individuals by race to only keep those of European descent. Next we removed all variants that are non-autosomal, as well as those that have at least one multi-character allele code or a nonstandard single-character allele code. Then we removed variants with missing genotype data and tested the remaining ones for Hardy-Weinberg equilibrium (HWE). In particular, we used the mid-p method to test for HWE (Wigginton, Cutler and Abecasis (2005); Graffelman and Moreno (2013)) and discarded SNPs with $P < 10^{-6}$. Lastly, we filtered out SNPs with minor allele frequency (MAF) below 0.01. After all of these steps, 715 individuals and 8,042,975 genomic sites remained. All genetic quality control steps were conducted in PLINK 1.90 (Chang et al. (2015); Purcell and Chang) and the resulting files were imported into R through the `BEDMatrix` package (Grueneberg and de los Campos (2019)).

DRAB supports covariate adjustment with any user-specified covariates. In this study we included the pre-computed European covariates that are available from the GTEx Portal. Namely, we downloaded and included the top 5 genotyping principal components, a set of latent covariates identified using the Probabilistic Estimation of Expression Residuals (PEER) method, and indicators for sequencing platform (Illumina HiSeq 2000 or HiSeq X), sequencing protocol (PCR-based or PCR-free), and sex. For details on the definitions of these covariates, see The GTEx Consortium (2020).

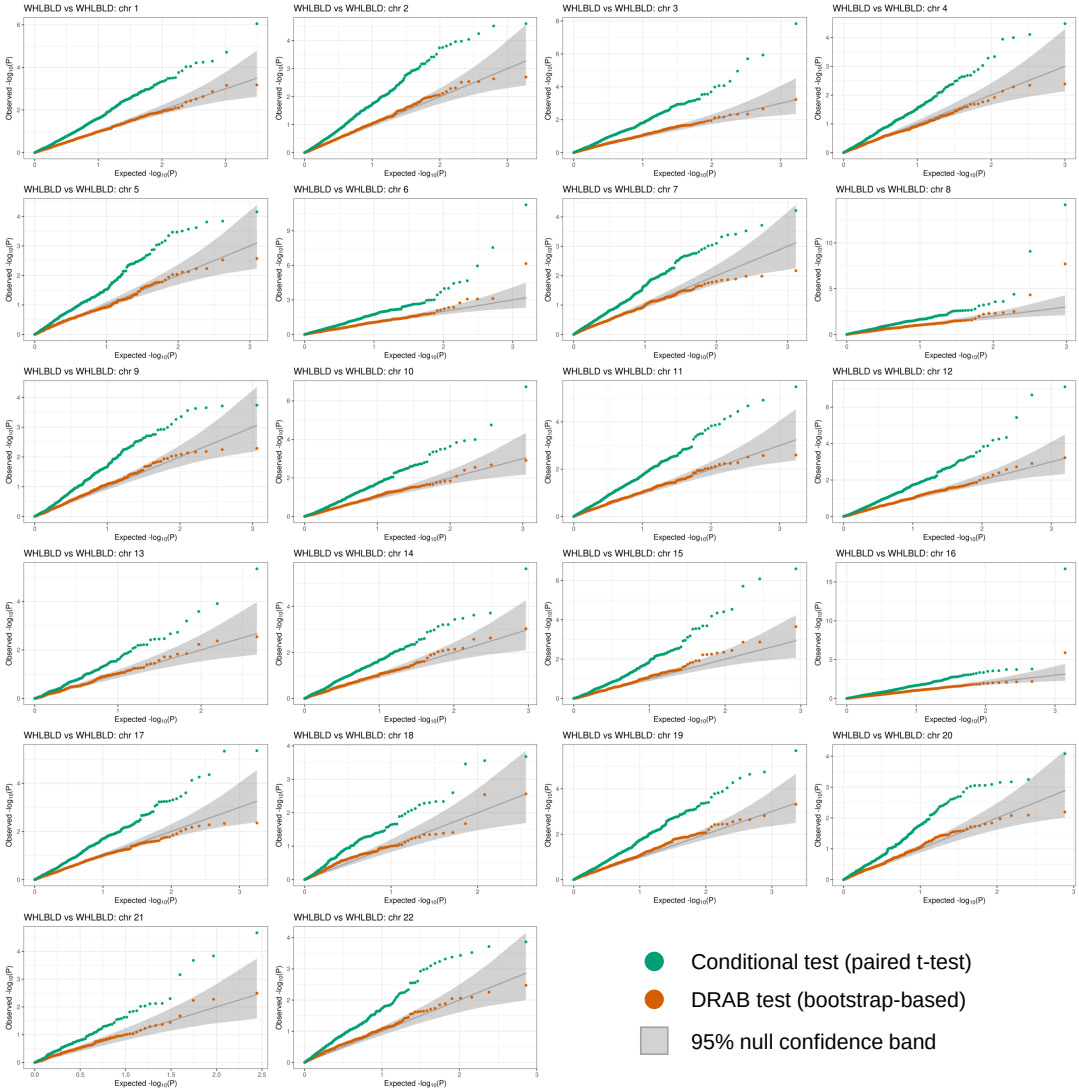

FIG 2. *Per-chromosome quantile-quantile (QQ) plots of P-values from differential regulation tests on whole blood (WHLBLD) data. The full data set for whole blood was partitioned into three equal-sized subsets, with two subsets used for training transcriptome prediction models and the third subset used for testing. All autosomal protein-coding genes were included in the analysis. Orange points denote $-\log_{10}(P)$ values from the bootstrap-based DRAB test, while green points denote $-\log_{10}(P)$ values from a paired t-test on the model prediction errors. The theoretical null distribution (assuming no subset-specific genetic regulation) is plotted as a solid gray line, flanked by 95% confidence bands.*

3.2. *The DRAB test accounts for model training uncertainty.* To demonstrate the advantages of our proposed bootstrap model comparison test over conditional tests, we evaluated the performance of both approaches on a single-tissue negative control. The negative control scenario was created by partitioning the set of individuals for whom gene expression measurements are available in whole blood tissue ($n = 558$) into three equal-sized subsets. Two of these subsets were treated as the training sets D_A and D_B, while the remaining one was treated as the test set D_T. Note that since the three subsets are nonoverlapping and have equal sizes, they satisfy the independence and cardinality assumptions stated in Section 2.2. We then applied DRAB to these data sets. Additionally, we created a conditional version of the DRAB framework by replacing our bootstrap model comparison test with a paired samples

t-test on the model prediction errors, as detailed in Section 2.4.1. The "conditional DRAB" test was also applied to the same data. We performed both of these differential regulation tests on each autosomal protein-coding gene, as defined by the GENCODE 26 transcript model provided on the GTEx Portal (The GTEx Consortium (2020)).

By construction, our negative control represents the scenario where context A and context B are sampled from the same population. Consequently, the patterns of genetic regulation are equivalent at the population level between the two contexts. If the context-specific models select different sets of eQTLs with differing effect sizes, then we know that those differences are only due to natural random variation and the instability of the learning algorithm. Ideally, a differential regulation analysis test should not identify any significant genes between the two subsets.

Figure 2 displays the per-chromosome results from our proposed DRAB test as well as from the analogous conditional test. The quantile-quantile (QQ) plots compare the observed distribution of P-values with their expected distribution under the null hypothesis of no differences in genetic regulatory patterns between the two subsets. Since we intentionally constructed the training data to ensure that the null hypothesis is true, the observed $-\log_{10}(P)$ values on each QQ plot should mostly lie within the 95% confidence band surrounding the identity line (shown in gray). However, this is clearly not the case for the conditional model comparison test. When the paired samples t-test is used to determine model equivalency, we observe extreme inflation of the $-\log_{10}(P)$ values. In fact, most of the P-values returned by the conditional test lie outside of the 95% confidence bands. Even after a Bonferroni correction for multiple testing, the conditional model comparison framework reported 19 genome-wide significant genes. Clearly, this is an unacceptably high type I error rate for ascertaining the presence of population-level differences in genetic regulation.

Our proposed bootstrap model comparison test, on the other hand, was able to effectively control the false positive rate. Notice that all of the DRAB P-values lie within the 95% confidence bands of Figure 2 (with only a couple minor exceptions, likely due to correlation between genes). Although the DRAB test is sometimes slightly conservative (e.g. on chromosome 17) or slightly liberal (e.g. on chromosome 15), overall it eliminated the hugely inflated type I error rates that were observed with a conditional test. This demonstrates that our proposed model comparison test is able to account for the variability of feature selection and model training, enabling rigorous comparisons of regression models at the population level.

Surprisingly, DRAB was able to effectively control false positives with relatively few bootstrap iterations. For all of the figures in this paper we used $K = 50$ bootstrap iterations, a much lower number than the hundreds or thousands of iterations usually recommended for the non-parametric bootstrap. This reduces the number of times that models need to be re-fit, making our framework computationally feasible to implement in real-world settings. Indeed, we found that DRAB yields acceptable results in the negative control scenario with as few as $K = 10$ bootstrap iterations, but it becomes somewhat unstable when the number of iterations is too low.

3.3. *Biologically related tissues share more regulatory patterns.* Having verified that DRAB is able to control the type I error rate, we next assessed its ability to detect tissue-specific patterns of genetic regulation between distinct tissues. We applied DRAB to 19 pairs comprised of 23 selected human tissues from the GTEx v8 data release. Since running DRAB on all possible pairings of the 49 tissues available in GTEx would have been computationally prohibitive, we aimed to select tissue pairs that represent a wide spectrum of tissues and tissue relationships. In particular, we made sure to include comparisons between whole blood and various nervous system tissues because whole blood is the most commonly used bulk

TABLE 1
Full list of considered tissues

Tissue site	Abbreviation	Samples
Adipose - Subcutaneous	ADPSBQ	479
Adrenal gland	ADRNLG	194
Artery - Aorta	ARTAORT	329
Brain - Amygdala	BRNAMY	119
Brain - Anterior cingulate cortex (BA24)	BRNACC	135
Brain - Cerebellum	BRNCHA	188
Brain - Cortex	BRNCTXA	183
Brain - Frontal cortex (BA9)	BRNCTXB	157
Brain - Hippocampus	BRNHPP	150
Brain - Spinal cord (cervical c-1)	BRNSPC	115
Brain - Substantia nigra	BRNSNG	100
Cells - EBV-transformed lymphocytes	LCL	113
Cells - Cultured fibroblasts	FIBRBLS	403
Colon - Sigmoid	CLNSGM	266
Liver	LIVER	178
Lung	LUNG	436
Muscle - Skeletal	MSCLSK	588
Nerve - Tibial	NERVET	438
Pancreas	PNCREAS	243
Skin - Sun exposed (lower leg)	SKINS	508
Small intestine - Terminal ileum	SNTTRM	141
Thyroid	THYROID	482
Whole blood	WHLBLD	558

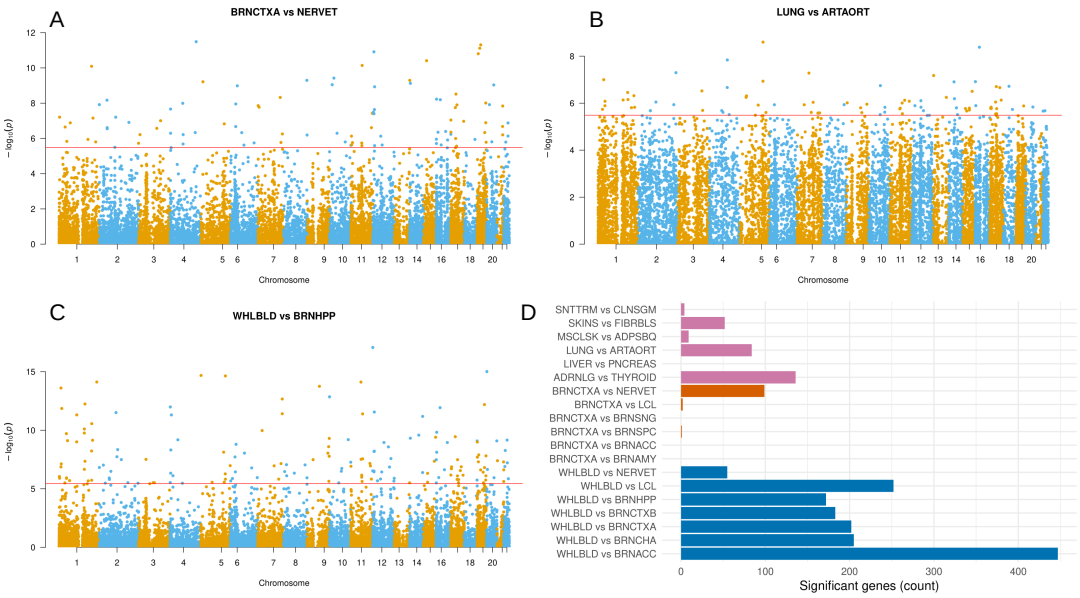

FIG 3. *(A) - (C) Manhattan plots of DRAB test results for pairings of various tissues. The horizontal red line denotes genome-wide significance according to a Bonferroni multiple testing correction. (D) Numbers of differentially regulated protein-coding genes for each of 19 different tissue pairs. Statistical significance was determined according to a genome-wide Bonferroni correction with a 95% confidence level.*

tissue in TWAS, owing to its relative ease of procurement. However, we suspected that blood might not be representative of genetic regulatory profiles in highly specialized brain tissues. We also tested for differential regulation between multiple brain regions. Finally, we chose six miscellaneous tissue pairs to further illustrate how extensively the genetic architecture of gene expression can vary between tissue types. The full list of considered tissues, along with their abbreviations and sample sizes, is displayed in Table 1.

Our results demonstrate that DRAB does have sufficient power to detect genes with tissue-specific patterns of genetic regulation. Figure 3 shows Manhattan plots of DRAB test results for all autosomal protein-coding genes in three tissue pairs. The tissues in each of these pairs have markedly different biological function and structure: for example, blood is entirely unlike the hippocampus. Therefore, DRAB should be able to identify a large number of differentially regulated genes in each of these tissue pairs. The DRAB test was indeed able to do so. After a Bonferroni correction for multiple testing, 99 genes were genome-wide significant between brain cortex and tibial nerve, 84 genes were genome-wide significant between lung and aorta, and 172 genes were genome-wide significant between whole blood and brain hippocampus. Figure 3 also displays a bar chart with the counts of significant protein-coding genes for the other tissue pairs that we considered. All gene-level P-values, as well as training and test sample sizes, are available in the Supplementary Material (Malakhov et al. (2023a)).

Notably, our results confirm that biologically related tissues have more similar patterns of genetic regulation. It has been experimentally established that differences in genetic regulation, and the subsequent over- or under-expression of certain genes, are the primary drivers of tissue differentiation in animals. Consequently, tissues with similar histological structure and function are expected to exhibit similar patterns of genetic regulation. The DRAB test concurs. As shown in Figure 3, pairs of closely related brain regions have few if any genes with significantly different regulatory profiles. When comparing blood and brain tissues, on the other hand, hundreds of genes display strong evidence of tissue-specific genetic regulation. This trend also holds for the miscellaneous tissue pairs that we considered, supporting the biological reasonableness of DRAB results.

4. Discussion. In this paper we have introduced DRAB (Differential Regulation Analysis by Bootstrapping) and made it available as open-source software both in the Supplementary Material (Malakhov et al. (2023b)) and on GitHub (`https://github.com/MykMal/drab`). Our method is a statistical framework for identifying genes that have significantly different patterns of local genetic regulation between two tissues or other biological contexts. DRAB utilizes penalized regression to learn context-specific models of the *cis*-genetic regulation of gene expression, and then applies a novel model comparison test to determine whether the genetic regulation models have equivalent predictive performance. The primary methodological innovations of our approach are twofold. First, we have proposed and confirmed that genetic regulatory profiles can be analyzed by training models to predict expression from eQTLs. Transcriptome imputation models are already widely used in statistical genetics, for instance as the first stage of TWAS, and our work illustrates another avenue for their application. Second, and perhaps more importantly, we have developed a new bootstrap model comparison test. The DRAB test incorporates both the variance associated with estimating the mean difference in model performance, as well as the variance associated with training the models themselves. Decomposing the variance in such a way makes it possible to compare the true, population-level models of genetic regulation despite the fact that they are inherently unknowable.

We have demonstrated that DRAB yields biologically reasonable results when applied to a variety of human tissue pairs from the GTEx Project. DRAB was able to effectively control

type I error when run on a negative control created from random samples of the same tissue. It also had sufficient power to detect genes that exhibit tissue-specific differences in regulatory profiles when such genes were expected to be found. Moreover, DRAB detected fewer differentially regulated genes in tissue pairs that share similar structure and function (e.g. different brain regions) than in pairs of biologically unrelated tissues (e.g. blood and brain tissues). This result is consistent with the contemporary understanding of gene expression and its relation to tissue function. We also showed that DRAB achieves good performance with as few as ten bootstrap iterations, making it computationally feasible to apply on genome-scale data from large consortia.

Although our application was only concerned with identifying genes that have tissue-specific patterns of genetic regulation of mRNA expression, the DRAB framework has broad relevance to many other problems in molecular genetics and applied statistics. In fact, DRAB can also test for differential genetic regulation in other gene expression products (i.e. other omic data types). The software package that we developed is directly applicable to any genetically influenced quantitative molecular phenotypes, including micro RNA expression, intron excision ratios, and measurements of transcription factor activity. More generally, our bootstrap model comparison test facilitates the rigorous assessment of predictive performance in any regression task where multiple competing parametric models are being considered. Model comparison and benchmarking are active areas of machine learning research, so we plan to further develop the ideas introduced herein.

We conclude with several limitations of our approach. First, we have not conducted a theoretical study of the proposed model comparison test; for asymptotic validity some conditions may be necessary, such as consistency and asymptotic normality of parameter estimation with the candidate models (Politis and Romano (1994)). Since we are considering regular parametric models and the elastic net is known to have model selection consistency (Yuan and Lin (2007); Jia and Yu (2010)), we expect those conditions to hold. However, it is possible that some adjustments might be needed, such as refitting selected models by least squares without penalization (Zhao, Witten and Shojaie (2021)), or de-biasing via some desparsified lasso and ridge projection procedures (Dezeure et al. (2015)). More generally, the proposed method may or may not perform well with non-parametric models due to technical challenges (Dai, Shen and Pan (2022)). Second, DRAB does not estimate a meaningful effect size for differential regulation. It can identify the genes that are differentially regulated between any two biological sample types, but it cannot determine whether some of those genes exhibit more extensive differences in genetic regulation than the others. For the same reason, DRAB is also unable to reconstruct the regulatory networks that underlie its results. We suggest possibly using DRAB in conjunction with differential network inference methods, which aim to build context-specific regulatory networks but cannot perform tests of statistical significance (Bhuva et al. (2019); Baur et al. (2020); Yazdani et al. (2020)). Third, the utility of DRAB ultimately depends on its ability to learn accurate models of transcriptional regulation. A similar limitation exists in TWAS, which is known to lose power when the R^2 of transcriptome imputation models is low (He et al. (2022)). In future work it may be beneficial to evaluate how the choice of prediction algorithm affects the power of DRAB to detect differentially regulated genes. Lastly, a major practical downside of our method is that it requires individual-level whole-genome genotypic and phenotypic data. Extensions of DRAB to genome-wide association study (GWAS) summary data would greatly expand the scope of its applications in genetics.

Acknowledgments. Access to the GTEx data was approved for dbGaP Project #26511, and the data were obtained from dbGaP accession number phs000424.v8.p2 on 10/13/2021. The authors also thank the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the results reported within this paper.

Funding. MMM was supported by NIH grant T32GM132063; XTS and WP were supported by NIH grants U01AG073079, R01AG065636, R01AG069895, R01AG074858, and R01GM126002. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary data for "A bootstrap model comparison test for identifying genes with context-specific patterns of genetic regulation"
This ZIP archive contains TSV files with all of the numerical results described in the paper. Included are the DRAB test P-values, conditional t-test P-values, and training/test set sample sizes for every tissue pair and gene.

Source code for "A bootstrap model comparison test for identifying genes with context-specific patterns of genetic regulation"
This ZIP archive contains all of the code necessary to reproduce our analyses, including an R implementation of DRAB and supporting bash scripts. For the most recent version of DRAB, see `https://github.com/MykMal/drab`.

REFERENCES

BARBEIRA, A. N., DICKINSON, S. P., BONAZZOLA, R., ZHENG, J., WHEELER, H. E., TORRES, J. M., TORSTENSON, E. S., SHAH, K. P., GARCIA, T., EDWARDS, T. L., STAHL, E. A., HUCKINS, L. M., GTEX CONSORTIUM, NICOLAE, D. L., COX, N. J. and IM, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9** 1825. https://doi.org/10.1038/s41467-018-03621-1

BARBEIRA, A. N., BONAZZOLA, R., GAMAZON, E. R., LIANG, Y., PARK, Y., KIM-HELLMUTH, S., WANG, G., JIANG, Z., ZHOU, D., HORMOZDIARI, F., LIU, B., RAO, A., HAMEL, A. R., PIVIDORI, M. D., AGUET, F., GTEX GWAS WORKING GROUP, BASTARACHE, L., JORDAN, D. M., VERBANCK, M., DO, R., GTEX CONSORTIUM, STEPHENS, M., ARDLIE, K., MCCARTHY, M., MONTGOMERY, S. B., SEGRÈ, A. V., BROWN, C. D., LAPPALAINEN, T., WEN, X. and IM, H. K. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biology* **22** 49. https://doi.org/10.1186/s13059-020-02252-4

BAUR, B., SHIN, J., ZHANG, S. and ROY, S. (2020). Data integration for inferring context-specific gene regulatory networks. *Current Opinion in Systems Biology* **23** 38–46. https://doi.org/10.1016/j.coisb.2020.09.005

BHUVA, D. D., CURSONS, J., SMYTH, G. K. and DAVIS, M. J. (2019). Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome Biology* **20** 236. https://doi.org/10.1186/s13059-019-1851-8

BULLARD, J. H., PURDOM, E., HANSEN, K. D. and DUDOIT, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11** 94. https://doi.org/10.1186/1471-2105-11-94

CADIOU, S. and SLAMA, R. (2021). Instability of Variable-selection Algorithms Used to Identify True Predictors of an Outcome in Intermediate-dimension Epidemiologic Studies. *Epidemiology* **32** 402–411. https://doi.org/10.1097/EDE.0000000000001340

CHANG, C. C., CHOW, C. C., TELLIER, L. C., VATTIKUTI, S., PURCELL, S. M. and LEE, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4** 7. https://doi.org/10.1186/s13742-015-0047-8

THE GTEX CONSORTIUM (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369** 1318–1330. https://doi.org/10.1126/science.aaz1776

DAI, B., SHEN, X. and PAN, W. (2022). Significance Tests of Feature Relevance for a Black-Box Learner. *IEEE Transactions on Neural Networks and Learning Systems* 1–14. https://doi.org/10.1109/TNNLS.2022.3185742

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statistical Science* **30** 533–558. https://doi.org/10.1214/15-STS527

EFRON, B. and TIBSHIRANI, R. (1993). *An introduction to the bootstrap*. *Monographs on statistics and applied probability* **57**. Chapman & Hall, New York.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**. https://doi.org/10.18637/jss.v033.i01

FRYETT, J. J., MORRIS, A. P. and CORDELL, H. J. (2020). Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. *Genetic Epidemiology* **44** 425–441. https://doi.org/10.1002/gepi.22290

GALLAGHER, M. D. and CHEN-PLOTKIN, A. S. (2018). The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics* **102** 717–730. https://doi.org/10.1016/j.ajhg.2018.04.002

GAMAZON, E. R., WHEELER, H. E., SHAH, K. P., MOZAFFARI, S. V., AQUINO-MICHAELS, K., CARROLL, R. J., EYLER, A. E., DENNY, J. C., GTEX CONSORTIUM, NICOLAE, D. L., COX, N. J. and IM, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47** 1091–1098. https://doi.org/10.1038/ng.3367

GILLIES, C. E., PUTLER, R., MENON, R., OTTO, E., YASUTAKE, K., NAIR, V., HOOVER, P., LIEB, D., LI, S., EDDY, S., FERMIN, D., MCNULTY, M. T., HACOHEN, N., KIRYLUK, K., KRETZLER, M., WEN, X., SAMPSON, M. G., SEDOR, J., DELL, K., SCHACHERE, M., LEMLEY, K., WHITTED, L., SRIVASTAVA, T., HANEY, C., SETHNA, C., GRAMMATIKOPOULOS, K., APPEL, G., TOLEDO, M., GREENBAUM, L., WANG, C.-S., LEE, B., ADLER, S., NAST, C., LAPAGE, J., ATHAVALE, A., NEU, A., BOYNTON, S., FERVENZA, F., HOGAN, M., LIESKE, J. C., CHERNITSKIY, V., KASKEL, F., KUMAR, N., FLYNN, P., KOPP, J., CASTRO-RUBIO, E., BLAKE, J., TRACHTMAN, H., ZHDANOVA, O., MODERSITZKI, F., VENTO, S., LAFAYETTE, R., MEHTA, K., GADEGBEKU, C., JOHNSTONE, D., CATTRAN, D., HLADUNEWICH, M., REICH, H., LING, P., ROMANO, M., FORNONI, A., BARISONI, L., BIDOT, C., KRETZLER, M., GIPSON, D., WILLIAMS, A., PITTER, R., NACHMAN, P., GIBSON, K., GRUBBS, S., FROMENT, A., HOLZMAN, L., MEYERS, K., KALLEM, K., CERECINO, F., SAMBANDAM, K., BROWN, E., JOHNSON, N., JEFFERSON, A., HINGORANI, S., TUTTLE, K., CURTIN, L., DISMUKE, S., COOPER, A., FREEDMAN, B., LIN, J. J., GRAY, S., KRETZLER, M., BARISONI, L., GADEGBEKU, C., GILLESPIE, B., GIPSON, D., HOLZMAN, L., MARIANI, L., SAMPSON, M. G., SONG, P., TROOST, J., ZEE, J., HERRESHOFF, E., KINCAID, C., LIENCZEWSKI, C., MAINIERI, T., WILLIAMS, A., ABBOTT, K., ROY, C., URV, T. and BROOKS, J. (2018). An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. *The American Journal of Human Genetics* **103** 232–244. https://doi.org/10.1016/j.ajhg.2018.07.004

GRAFFELMAN, J. and MORENO, V. (2013). The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Statistical Applications in Genetics and Molecular Biology* **12**. https://doi.org/10.1515/sagmb-2012-0039

GRUENEBERG, A. and DE LOS CAMPOS, G. (2019). BGData - A Suite of R Packages for Genomic Analysis with Big Data. *G3 Genes|Genomes|Genetics* **9** 1377–1383. https://doi.org/10.1534/g3.119.400018

GUO, X., LIN, W., WEN, W., HUYGHE, J., BIEN, S., CAI, Q., HARRISON, T., CHEN, Z., QU, C., BAO, J., LONG, J., YUAN, Y., WANG, F., BAI, M., ABECASIS, G. R., ALBANES, D., BERNDT, S. I., BÉZIEAU, S., BISHOP, D. T., BRENNER, H., BUCH, S., BURNETT-HARTMAN, A., CAMPBELL, P. T., CASTELLVÍ-BEL, S., CHAN, A. T., CHANG-CLAUDE, J., CHANOCK, S. J., CHO, S. H., CONTI, D. V., CHAPELLE, A. D. L., FESKENS, E. J. M., GALLINGER, S. J., GILES, G. G., GOODMAN, P. J., GSUR, A., GUINTER, M., GUNTER, M. J., HAMPE, J., HAMPEL, H., HAYES, R. B., HOFFMEISTER, M., KAMPMAN, E., KANG, H. M., KEKU, T. O., KIM, H. R., LE MARCHAND, L., LEE, S. C., LI, C. I., LI, L., LINDBLOM, A., LINDOR, N., MILNE, R. L., MORENO, V., MURPHY, N., NEWCOMB, P. A., NICKERSON, D. A., OFFIT, K., PEARLMAN, R., PHAROAH, P. D. P., PLATZ, E. A., POTTER, J. D., RENNERT, G., SAKODA, L. C., SCHAFMAYER, C., SCHMIT, S. L., SCHOEN, R. E., SCHUMACHER, F. R., SLATTERY, M. L., SU, Y.-R., TANGEN, C. M., ULRICH, C. M., VAN DUIJNHOVEN, F. J. B., VAN GUELPEN, B., VISVANATHAN, K., VODICKA, P., VODICKOVA, L., VYMETALKOVA, V., WANG, X., WHITE, E., WOLK, A., WOODS, M. O., CASEY, G., HSU, L., JENKINS, M. A., GRUBER, S. B., PETERS, U. and ZHENG, W. (2021). Identifying Novel Susceptibility Genes for Colorectal Cancer Risk From a Transcriptome-Wide Association Study of 125,478 Subjects. *Gastroenterology* **160** 1164–1178.e6. https://doi.org/10.1053/j.gastro.2020.08.062

GUSEV, A., KO, A., SHI, H., BHATIA, G., CHUNG, W., PENNINX, B. W. J. H., JANSEN, R., DE GEUS, E. J. C., BOOMSMA, D. I., WRIGHT, F. A., SULLIVAN, P. F., NIKKOLA, E., ALVAREZ, M., CIVELEK, M., LUSIS, A. J., LEHTIMÄKI, T., RAITOHARJU, E., KÄHÖNEN, M., SEPPÄLÄ, I., RAITAKARI, O. T., KUUSISTO, J., LAAKSO, M., PRICE, A. L., PAJUKANTA, P. and PASANIUC, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48** 245–252. https://doi.org/10.1038/ng.3506

HE, R., XUE, H., PAN, W. and FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2022). Statistical power of transcriptome-wide association studies. *Genetic Epidemiology* **46** 572–588. https://doi.org/10.1002/gepi.22491

HOERL, A. E. and KENNARD, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12** 55–67. https://doi.org/10.1080/00401706.1970.10488634

HU, Y., LI, M., LU, Q., WENG, H., WANG, J., ZEKAVAT, S. M., YU, Z., LI, B., GU, J., MUCHNIK, S., SHI, Y., KUNKLE, B. W., MUKHERJEE, S., NATARAJAN, P., NAJ, A., KUZMA, A., ZHAO, Y.,

CRANE, P. K., ALZHEIMER'S DISEASE GENETICS CONSORTIUM, LU, H. and ZHAO, H. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics* **51** 568–576. https://doi.org/10.1038/s41588-019-0345-7

JIA, J. and YU, B. (2010). On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica* **20** 595–611.

KALOUSIS, A., PRADOS, J. and HILARIO, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems* **12** 95–116. https://doi.org/10.1007/s10115-006-0040-8

LIN, Z., XUE, H., MALAKHOV, M. M., KNUTSON, K. A. and PAN, W. (2022). Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptome-wide association studies. *Human Molecular Genetics* **31** 2462–2470. https://doi.org/10.1093/hmg/ddac015

LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F., YOUNG, N., FOSTER, B., MOSER, M., KARASIK, E., GILLARD, B., RAMSEY, K., SULLIVAN, S., BRIDGE, J., MAGAZINE, H., SYRON, J., FLEMING, J., SIMINOFF, L., TRAINO, H., MOSAVEL, M., BARKER, L., JEWELL, S., ROHRER, D., MAXIM, D., FILKINS, D., HARBACH, P., CORTADILLO, E., BERGHUIS, B., TURNER, L., HUDSON, E., FEENSTRA, K., SOBIN, L., ROBB, J., BRANTON, P., KORZENIEWSKI, G., SHIVE, C., TABOR, D., QI, L., GROCH, K., NAMPALLY, S., BUIA, S., ZIMMERMAN, A., SMITH, A., BURGES, R., ROBINSON, K., VALENTINO, K., BRADBURY, D., COSENTINO, M., DIAZ-MAYORAL, N., KENNEDY, M., ENGEL, T., WILLIAMS, P., ERICKSON, K., ARDLIE, K., WINCKLER, W., GETZ, G., DELUCA, D., MACARTHUR, D., KELLIS, M., THOMSON, A., YOUNG, T., GELFAND, E., DONOVAN, M., MENG, Y., GRANT, G., MASH, D., MARCUS, Y., BASILE, M., LIU, J., ZHU, J., TU, Z., COX, N. J., NICOLAE, D. L., GAMAZON, E. R., IM, H. K., KONKASHBAEV, A., PRITCHARD, J., STEVENS, M., FLUTRE, T., WEN, X., DERMITZAKIS, E. T., LAPPALAINEN, T., GUIGO, R., MONLONG, J., SAMMETH, M., KOLLER, D., BATTLE, A., MOSTAFAVI, S., MCCARTHY, M., RIVAS, M., MALLER, J., RUSYN, I., NOBEL, A., WRIGHT, F., SHABALIN, A., FEOLO, M., SHAROPOVA, N., STURCKE, A., PASCHAL, J., ANDERSON, J. M., WILDER, E. L., DERR, L. K., GREEN, E. D., STRUEWING, J. P., TEMPLE, G., VOLPI, S., BOYER, J. T., THOMSON, E. J., GUYER, M. S., NG, C., ABDALLAH, A., COLANTUONI, D., INSEL, T. R., KOESTER, S. E., LITTLE, A. R., BENDER, P. K., LEHNER, T., YAO, Y., COMPTON, C. C., VAUGHT, J. B., SAWYER, S., LOCKHART, N. C., DEMCHOK, J. and MOORE, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45** 580–585. https://doi.org/10.1038/ng.2653

MALAKHOV, M. M., DAI, B., SHEN, X. T. and PAN, W. (2023a). Supplementary data for "A bootstrap model comparison test for identifying genes with context-specific patterns of genetic regulation".

MALAKHOV, M. M., DAI, B., SHEN, X. T. and PAN, W. (2023b). Source code for "A bootstrap model comparison test for identifying genes with context-specific patterns of genetic regulation".

NOGUEIRA, S., SECHIDIS, K. and BROWN, G. (2018). On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research* **18** 1–54.

OKORO, P. C., SCHUBERT, R., GUO, X., JOHNSON, W. C., ROTTER, J. I., HOESCHELE, I., LIU, Y., IM, H. K., LUKE, A., DUGAS, L. R. and WHEELER, H. E. (2021). Transcriptome prediction performance across machine learning models and diverse ancestries. *Human Genetics and Genomics Advances* **2** 100019. https://doi.org/10.1016/j.xhgg.2020.100019

PAN, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18** 546–554. https://doi.org/10.1093/bioinformatics/18.4.546

POLITIS, D. N. and ROMANO, J. P. (1994). Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions. *The Annals of Statistics* **22** 2031–2050. https://doi.org/10.1214/aos/1176325770

PURCELL, S. and CHANG, C. PLINK 1.90. Version beta 7, 64-bit (16 Jan 2023).

SEYEDNASROLLAH, F., LAIHO, A. and ELO, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics* **16** 59–70. https://doi.org/10.1093/bib/bbt086

SONESON, C. and DELORENZI, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14** 91. https://doi.org/10.1186/1471-2105-14-91

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J. and TIBSHIRANI, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 245–266. https://doi.org/10.1111/j.1467-9868.2011.01004.x

WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences* **117** 16880–16890. https://doi.org/10.1073/pnas.1922664117

WHEELER, H. E., SHAH, K. P., BRENNER, J., GARCIA, T., AQUINO-MICHAELS, K., GTEX CONSORTIUM, COX, N. J., NICOLAE, D. L. and IM, H. K. (2016). Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *PLOS Genetics* **12** e1006423. https://doi.org/10.1371/journal.pgen.1006423

WIGGINTON, J. E., CUTLER, D. J. and ABECASIS, G. R. (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics* **76** 887–893. https://doi.org/10.1086/429864

WONG, A. K., SEALFON, R. S. G., THEESFELD, C. L. and TROYANSKAYA, O. G. (2021). Decoding disease: from genomes to networks to phenotypes. *Nature Reviews Genetics* **22** 774–790. https://doi.org/10.1038/s41576-021-00389-x

WU, L., SHI, W., LONG, J., GUO, X., MICHAILIDOU, K., BEESLEY, J., BOLLA, M. K., SHU, X.-O., LU, Y., CAI, Q., AL-EJEH, F., ROZALI, E., WANG, Q., DENNIS, J., LI, B., ZENG, C., FENG, H., GUSEV, A., BARFIELD, R. T., ANDRULIS, I. L., ANTON-CULVER, H., ARNDT, V., ARONSON, K. J., AUER, P. L., BARRDAHL, M., BAYNES, C., BECKMANN, M. W., BENITEZ, J., BERMISHEVA, M., BLOMQVIST, C., BOGDANOVA, N. V., BOJESEN, S. E., BRAUCH, H., BRENNER, H., BRINTON, L., BROBERG, P., BRUCKER, S. Y., BURWINKEL, B., CALDÉS, T., CANZIAN, F., CARTER, B. D., CASTELAO, J. E., CHANG-CLAUDE, J., CHEN, X., CHENG, T.-Y. D., CHRISTIANSEN, H., CLARKE, C. L., NBCS COLLABORATORS, COLLÉE, M., CORNELISSEN, S., COUCH, F. J., COX, D., COX, A., CROSS, S. S., CUNNINGHAM, J. M., CZENE, K., DALY, M. B., DEVILEE, P., DOHENY, K. F., DÖRK, T., DOS SANTOS-SILVA, I., DUMONT, M., DWEK, M., ECCLES, D. M., EILBER, U., ELIASSEN, A. H., ENGEL, C., ERIKSSON, M., FACHAL, L., FASCHING, P. A., FIGUEROA, J., FLESCH-JANYS, D., FLETCHER, O., FLYGER, H., FRITSCHI, L., GABRIELSON, M., GAGO-DOMINGUEZ, M., GAPSTUR, S. M., GARCÍA-CLOSAS, M., GAUDET, M. M., GHOUSSAINI, M., GILES, G. G., GOLDBERG, M. S., GOLDGAR, D. E., GONZÁLEZ-NEIRA, A., GUÉNEL, P., HAHNEN, E., HAIMAN, C. A., HÅKANSSON, N., HALL, P., HALLBERG, E., HAMANN, U., HARRINGTON, P., HEIN, A., HICKS, B., HILLEMANNS, P., HOLLESTELLE, A., HOOVER, R. N., HOPPER, J. L., HUANG, G., HUMPHREYS, K., HUNTER, D. J., JAKUBOWSKA, A., JANNI, W., JOHN, E. M., JOHNSON, N., JONES, K., JONES, M. E., JUNG, A., KAAKS, R., KERIN, M. J., KHUSNUTDINOVA, E., KOSMA, V.-M., KRISTENSEN, V. N., LAMBRECHTS, D., LE MARCHAND, L., LI, J., LINDSTRÖM, S., LISSOWSKA, J., LO, W.-Y., LOIBL, S., LUBINSKI, J., LUCCARINI, C., LUX, M. P., MACINNIS, R. J., MAISHMAN, T., KOSTOVSKA, I. M., MANNERMAA, A., MANSON, J. E., MARGOLIN, S., MAVROUDIS, D., MEIJERS-HEIJBOER, H., MEINDL, A., MENON, U., MEYER, J., MULLIGAN, A. M., NEUHAUSEN, S. L., NEVANLINNA, H., NEVEN, P., NIELSEN, S. F., NORDESTGAARD, B. G., OLOPADE, O. I., OLSON, J. E., OLSSON, H., PETERLONGO, P., PETO, J., PLASESKA-KARANFILSKA, D., PRENTICE, R., PRESNEAU, N., PYLKÄS, K., RACK, B., RADICE, P., RAHMAN, N., RENNERT, G., RENNERT, H. S., RHENIUS, V., ROMERO, A., ROMM, J., RUDOLPH, A., SALOUSTROS, E., SANDLER, D. P., SAWYER, E. J., SCHMIDT, M. K., SCHMUTZLER, R. K., SCHNEEWEISS, A., SCOTT, R. J., SCOTT, C. G., SEAL, S., SHAH, M., SHRUBSOLE, M. J., SMEETS, A., SOUTHEY, M. C., SPINELLI, J. J., STONE, J., SUROWY, H., SWERDLOW, A. J., TAMIMI, R. M., TAPPER, W., TAYLOR, J. A., TERRY, M. B., TESSIER, D. C., THOMAS, A., THÖNE, K., TOLLENAAR, R. A. E. M., TORRES, D., TRUONG, T., UNTCH, M., VACHON, C., VAN DEN BERG, D., VINCENT, D., WAISFISZ, Q., WEINBERG, C. R., WENDT, C., WHITTEMORE, A. S., WILDIERS, H., WILLETT, W. C., WINQVIST, R., WOLK, A., XIA, L., YANG, X. R., ZIOGAS, A., ZIV, E., KCONFAB/AOCS INVESTIGATORS, DUNNING, A. M., PHAROAH, P. D. P., SIMARD, J., MILNE, R. L., EDWARDS, S. L., KRAFT, P., EASTON, D. F., CHENEVIX-TRENCH, G. and ZHENG, W. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature Genetics* **50** 968–978. https://doi.org/10.1038/s41588-018-0132-x

YANG, T., WU, C., WEI, P. and PAN, W. (2020). Integrating DNA sequencing and transcriptomic data for association analyses of low-frequency variants and lipid traits. *Human Molecular Genetics* **29** 515–526. https://doi.org/10.1093/hmg/ddz314

YAZDANI, A., MENDEZ-GIRALDEZ, R., YAZDANI, A., KOSOROK, M. R. and ROUSSOS, P. (2020). Differential gene regulatory pattern in the human brain from schizophrenia using transcriptomic-causal network. *BMC Bioinformatics* **21** 469. https://doi.org/10.1186/s12859-020-03753-6

YUAN, M. and LIN, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 143–161. https://doi.org/10.1111/j.1467-9868.2007.00581.x

ZHAO, S., WITTEN, D. and SHOJAIE, A. (2021). In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference. *Statistical Science* **36** 562–577. https://doi.org/10.1214/20-STS815

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x