RESEARCH ARTICLE

Variable selection in omics data: A practical evaluation of small sample sizes

Alexander Kirpich[1,2], **Elizabeth A. Ainsworth**[3,4], **Jessica M. Wedow**[3], **Jeremy R. B. Newman**[1], **George Michailidis**[2,5], **Lauren M. McIntyre**[1,2,6]*

1 Department of Biology, University of Florida, Gainesville, FL, United States of America, **2** Informatics Institute, University of Florida, Gainesville, FL, United States of America, **3** Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, United States of America, **4** USDA ARS Global Change and Photosynthesis Research Unit, Urbana, IL, United States of America, **5** Department of Statistics, University of Florida, Gainesville, FL, United States of America, **6** Genetics Institute, University of Florida, Gainesville, FL, United States of America

* mcintyre@ufl.edu

Citation: Kirpich A, Ainsworth EA, Wedow JM, Newman JRB, Michailidis G, McIntyre LM (2018) Variable selection in omics data: A practical evaluation of small sample sizes. PLoS ONE 13(6): e0197910. https://doi.org/10.1371/journal.pone.0197910

Editor: Mohammed S. Orloff, University of Arkansas for Medical Sciences, UNITED STATES

Received: September 30, 2017

Accepted: May 10, 2018

Published: June 21, 2018

Data Availability Statement: Data are available from dbGaP under accession number phs000911.v1.p1 RNA-Seq data and related genomic data are both available from this source. RNAseq data is also available from dbGaP under accession number phs001426.v1.p1.

Funding: This work was funded by multiple sources that authors want to acknowledge: National Institutes of Health (NIH) U24 DK097209, National Science Foundation (NSF) 1238030, National Institute of Diabetes and Digestive and

Abstract

In omics experiments, variable selection involves a large number of metabolites/ genes and a small number of samples (the $n < p$ problem). The ultimate goal is often the identification of one, or a few features that are different among conditions- a biomarker. Complicating biomarker identification, the p variables often contain a correlation structure due to the biology of the experiment making identifying causal compounds from correlated compounds difficult. Additionally, there may be elements in the experimental design (blocks, batches) that introduce structure in the data. While this problem has been discussed in the literature and various strategies proposed, the over fitting problems concomitant with such approaches are rarely acknowledged. Instead of viewing a single omics experiment as a definitive test for a biomarker, an unrealistic analytical goal, we propose to view such studies as screening studies where the goal of the study is to reduce the number of features present in the second round of testing, and to limit the Type II error. Using this perspective, the performance of LASSO, ridge regression and Elastic Net was compared with the performance of an ANOVA via a simulation study and two real data comparisons. Interestingly, a dramatic increase in the number of features had no effect on Type I error for the ANOVA approach. ANOVA, even without multiple test correction, has a low false positive rates in the scenarios tested. The Elastic Net has an inflated Type I error (from 10 to 50%) for small numbers of features which increases with sample size. The Type II error rate for the ANOVA is comparable or lower than that for the Elastic Net leading us to conclude that an ANOVA is an effective analytical tool for the initial screening of features in omics experiments.

Introduction

In omics experiments, an analysis goal is the identification of features (metabolites or genes) that are different between treatment groups. The inspiration for this work is the analysis of untargeted metabolomics data, however the general results can be applied to other types of

Kidney Diseases (NIDDK) DP3DK085678, the University of Florida Informatics Institute (UFII) Fellowship, and the University of Florida Genetics Institute.

Competing interests: The authors have declared that no competing interests exist.

omics studies. In a typical metabolomics experiment the total number of samples is small and often limited by cost and processing time. The data collected are in the form of chromatograms, which are processed to select regions (peaks) that represent underlying chemical structures. Processing can be executed by proprietary instrument-specific software or with third party open source software, such as MZmine [1]. Selected peaks are quantified for each sample. This process has been well reviewed and well described [2] [3]. Peaks can be matched to metabolites/lipids and this identification process is its own area of research [4] [5]. In this work it is assumed, that quantified peaks (or gene expression levels) are the starting point for the analysis.

Typically, in omics experiments, the number of features is large in comparison to the number of samples and exceeds the number of available samples by orders of magnitude. Two alternative analysis paradigms are compared in this work. The first paradigm is based on classification approaches and compares the least absolute shrinkage and selection operator (LASSO), ridge regression and the generalization of these approaches—the Elastic Net feature selection methods, which accounts for the correlation structure among features. The second paradigm uses a linear models framework, where individual features are modeled separately ignoring the correlation structure among features, but allowing the incorporation of the experimental design structure directly into the model. By focusing on the "small" sample size *n* and "large" number of features *p* scenarios, these approaches are evaluated in settings likely to mimic those of the typical practitioner and on different datasets. Type I error and power have been estimated for the simulated data, to allow the practitioner to understand the performance of these approaches in real world settings.

Classification approaches have been deployed on subsets of original features selected after an initial "filtering" step. Prior comparisons deployed pre-screening of the original data, so that the final set of features available for further use and analysis was (much) smaller than the original set obtained from the instrument [6] [7] [8] [9] [10] [11]. Pre-screening using a *t*-test [6] [8] *t*-statistic scores [7] [12], Hardy-Weinberg equilibrium tests [9] and also non-statistical biological considerations [10] [6] have been proposed, with the subsequent application of statistical or machine learning methods to a subset of features. The rationale for the pre-screening of features is to aid in the efficient classification of samples into groups rather than feature selection [7] [13]. For example [6] after data pre-processing and pre-screening only 163 features remain out of $\sim$16,000 original data features. In the context of biomarker identification, feature selection rather than classification of samples is the goal. In the context of biomarker identification, the Type I and Type II errors of the entire process (pre-screening plus selection by machine learning) are unknown for real data. Here we focus on determining the Type I and Type II errors associated with feature selection using a single step.

In addition to the above studies which explicitly discuss "pre-screening" many methods developed for feature selection compare modern approaches using real data. In these comparisons, the prediction accuracy is used as the measure of performance since the Type I and Type II error are unknown in this setting. In a comparison of Multivariate Adaptive Regression Splines, Learning Ensemble (including bagging and boosting), Random Forest, Bayesian Moving Averaging, Stochastic Search Variable Selection, and Generalized Regularized Logistics Regression. The generalized regularized regression model (Elastic Net) had the highest predictive power [14]. A comparison of classifiers from 115 datasets found that Elastic Net was not different from bagging of *k*-nearest neighbors [15], support vector machine [16], and a 1-hidden layer neural network with sigmoid transfer function [17], [18]. In a recent comparison [19] of Boruta [20], the Vita method [21], recurrent relative variable importance [22], a parametric permutation approach [23] as well as recursive feature elimination (RFE) only permutation importance had some control of the Type I error. In direct comparisons of permutation importance and Elastic Net Random Forest was outperformed by Elastic Net [24].

In a simulation study performed by Acharjee et.al. [6] 100 samples were considered with 12 significant features out of 1000 comparing the performance of LASSO, Elastic Net, ridge regression, principal components regression (PCR), and other methods used for feature selection. The Elastic Net had the lowest mean squared error of prediction (MSEP) among the considered methods. LASSO, classification tree (CT), and linear discriminant analysis (LDA) were applied [8] for metabolic biosignature for Lyme disease prediction, where sample sizes were 202 and 259 for the treatment and control group, respectively. The number of features before and after pre-screening were 2262 and 95, respectively. Subgroups with sample sizes as small as 20 were evaluated. LASSO performed the best in terms of the receiver operating characteristic (ROC) curves for these data. Elastic Net had lower MSEP in comparison to support vector machines (SVM) and penalized logistic regression (PLSR) [7] [25]. LASSO has been shown to outperform OPLS-DA in feature selection [26] and Elastic Net is superior to stepwise selection [27].

LASSO, ridge regression and Elastic Net form a special class of penalized regression models. The first modeling approach with the penalty of that kind was ridge regression proposed by Hoerl and Kennard [28]. Due to the structure of the penalty ridge regression has a closed form solution for the standard linear models with normal errors and results in shrunk regression coefficients none of which is equal to zero. Thus, ridge regression can be used as a prediction tool but not as a feature selector directly. LASSO was introduced by Tibshirani [29]. While the LASSO method does not allow a closed form solutions it allows for variable selection. LASSO uses shrinkage to estimate which set of the regression coefficients have a value of zero and can therefore be eliminated. The LASSO method has limitations such as the number of variables that can be selected by the method has to be smaller or equal to the sample size n. The LASSO will also often select only a single feature in a set of highly correlated features [7].

The Elastic Net method introduced by Zou and Hastie [7] addressed the drawbacks of the LASSO and ridge regression methods, by creating a general framework and incorporated these two methods as special cases. The Elastic Net is a weighted combination of both LASSO and ridge regression penalties. The split between the penalties is controlled by the penalty split parameter $\alpha \in [0; 1]$ where $\alpha = 0$ corresponds to ridge regression penalty and $\alpha = 1$ corresponds to LASSO. Elastic Net was initially introduced for linear models with normal distribution of errors [7] and has later been extended to other types of models that included generalized linear models (such as logistic regression) and survival models together with efficient numerical computation algorithms [30] [31]. Elastic Net has also been shown to be a generalization of Support Vector Machines (SVM) [32] enabling some fast computational solutions developed for SVM to be applied to the Elastic Net. Elastic Net has been applied to the analysis of multiple real and simulated datasets, and has more than 2500 citations as of the writing of this manuscript. The computation algorithm proposed by Friedman et.al. [31] and the corresponding software package [33] has been used in this work.

For omics data the number of samples n is often between 10 and 100. Such small sample sizes represent a challenge for the applied statistician. Despite this, little attention has been given to very small sample sizes. For example [31] the dataset with the smallest sample size considered was a leukemia dataset with $n = 72$ samples and $p = 3571$ features. The data were originally collected by Golub et.al. [34]. The authors report that $p = 72$ features were selected by LASSO, all features ($p = 3571$) were selected by ridge regression, and the Elastic Net with $\alpha = 0.2$ selected values in between (Fig 1 [31]). In their simulation the smallest sample sizes considered was $n = 100$ [31]. Another simulation study comparing performance of ridge regression, LASSO and Elastic Net for small n and large p omics settings was conducted [35]. The

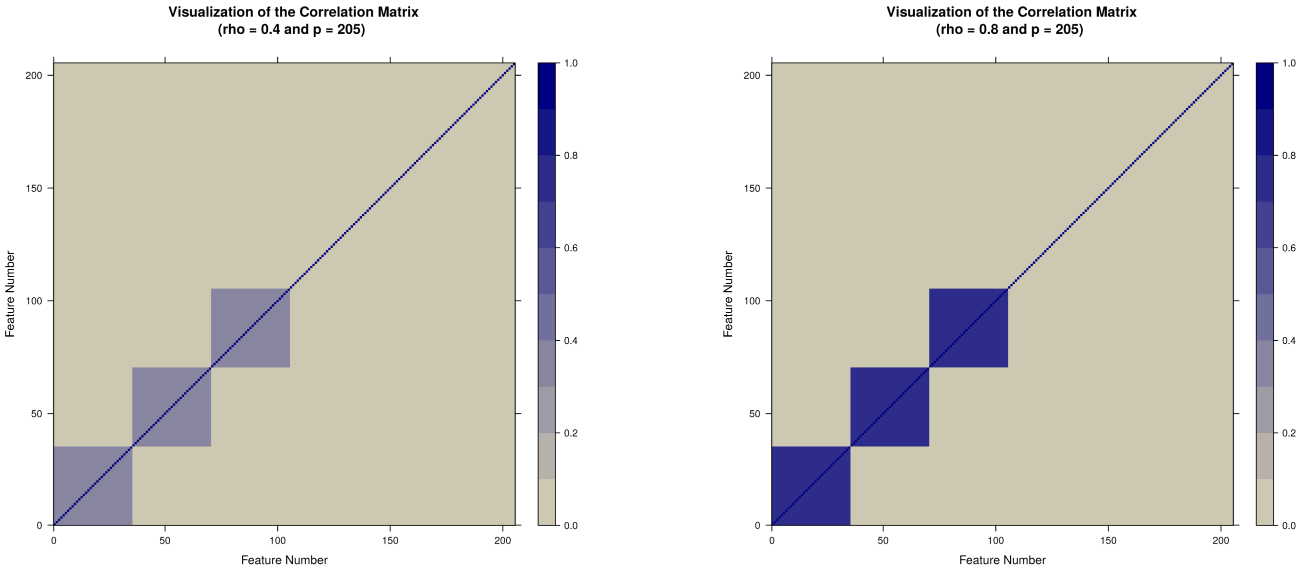

Fig 1. Visualization of power (left column) and Type I error (right column) estimates comparison between $p = 205$ (solid lines) and $p = 2050$ (dashed line) features for $\rho = 0.4$ and sample sizes $n = 10$ (top row), $n = 50$ (middle row), and $n = 100$ (bottom row). The value of the penalty split parameter α is plotted on the x-axis. Type I error and power estimates are plotted on y-axis for the values of α in the range of [0; 1] with 0.1 increments. In the left column power estimates are provided based on the four different features for each of the effect sizes ($\Delta_1 = 0.2$ is the red line, $\Delta_2 = 0.5$ is the blue line, and $\Delta_3 = 0.8$ is the green line). In the right column Type I error estimates are provided (beige lines) based on the random noise features together with a 0.05 threshold plotted as a purple dashed line. The vertical dashed line in the right column plots corresponds to penalty split value $\alpha = 0.5$. The value of $\alpha = 0$ corresponds to ridge regression and $\alpha = 1$ corresponds to LASSO.

https://doi.org/10.1371/journal.pone.0197910.g001

authors examined sample sizes in the range $n = 100, \ldots, 1000$ with a focus on prediction rather than feature selection. Ridge regression and Elastic Net were determined to outperform LASSO for prediction but the samples sizes smaller than 100 were not considered. Direct comparisons of LASSO, Elastic Net, classification and regression tree (CRT), random forest (RF), and t-tests for binary outcome and "non-big" data ($n > p$) using the simulation studies and psychiatric disorders data were performed [24]. The authors used sample sizes $n = 200, 400$ and $p = 60$ features for simulations which corresponds to the case of "non-big" data ($n > p$). For the real data considered in the paper the number of samples was $n = 475$ and the number of considered features was $p = 44$. The authors concluded that LASSO and Elastic Net provide superior performance for feature identification compared to other considered methods, and that a two-sided t-test performed well compared to LASSO and Elastic Net under certain scenarios [24].

In this work Elastic Net, LASSO and ridge regression are compared to an ANOVA approach for feature selection purposes with a focus on sample sizes less than 100. Simulation and two real data examples representing different aspects of omics studies are considered. In the first real dataset, a field experiment in maize, 24, B73 plants are evaluated in two conditions (ambient (n = 12) and elevated (n = 12) ozone). In the second experiment, 81 type 1 diabetics are examined in for gene expression differences in two different blood cell types (CD4+ and CD8+). In the maize experiment we expect very few differences among metabolites while in the diabetes data we expect large differences among the cell types. The goal in these and other similar experiments is to generate a list of features different between two experimental conditions for further evaluation; it is not to immediately generate a biomarker.

Materials and methods

Models

Given feature values Y_{ijk}, where i is an indicator of the feature ($i = 1, 2, \ldots, p$), j is an indicator of the sample ($j = 1, 2, \ldots, n$), and k is an indicator of the treatment group ($k = 1, 2, \ldots, m$), the primary interest is the identification of i-s such that corresponding features have different means for different treatment groups; i.e. those features where $\mu_{ik} \neq \mu_{ik'}$ for $k \neq k'$.

In a linear model the Y_{ijk} are dependent outcomes. Without loss of generality the case of two groups ($m = 2$) is considered in this work. The formal model for feature i is either formulated as means model:

$$\begin{gathered} Y_{ijk} = \mu_{ik} + \epsilon_{ijk} \text{ where } \epsilon_{ijk} \sim iid\, N(0, \tau^2) \\ j = 1, 2, \ldots, n \quad k = 1, 2, \end{gathered} \tag{1}$$

where μ_{ik} is the mean for feature i and group mean k. The alternative model formulation for feature i is the effects model:

$$\begin{gathered} Y_{ijk} = \mu_i + \nu_{ik} + \epsilon_{ijk} \text{ where } \epsilon_{ijk} \sim iid\, N(0, \tau^2) \\ j = 1, 2, \ldots, n \quad k = 1, 2, \end{gathered} \tag{2}$$

where μ_i is the overall mean for feature i and ν_{ik} is the mean for feature i and group k. The required identifiability constraint for the effects model is $\sum_{k=1}^{2} \nu_{ik} = 0$. The tests of hypothesis H_0: $\mu_{i1} = \mu_{i2}$ (or H_0: $\nu_{i1} = \nu_{i2}$) for each i with correlation among outcomes and a necessary adjustment for multiple testing to select features are used for comparison. The significance of difference for each feature i is determined by the corresponding ANOVA F-test. The ANOVA-based analysis was performed using `R` language and corresponding build-in function `aov`.

In the classification approach features $\boldsymbol{Y}_j = (Y_{1j}, Y_{2j}, \ldots, Y_{pj})^T$ are treated as predictors for the dependent binary random variable K_j that defines the group that sample j belongs to. The approach utilizes logistic regression with Elastic Net penalty proposed by Zou [7] and estimation algorithm proposed by Friedman et.al. [31]. The approach has been implemented in `R` package `glmnet` [31]. The detailed tutorial of the package has been provided online [36].

Let G_k represent the set of indexes of those samples that belong to group k and $I_{jk} = I(j \in G_k)$ be an indicator of sample j belonging to group k. For the two groups scenario the corresponding probabilities for sample j are modeled as

$$\begin{aligned} \Pr(K_j = 2 | \boldsymbol{Y}_j) &= \frac{\exp[\beta_0 + \boldsymbol{\beta}^T \boldsymbol{Y}_j]}{1 + \exp[\beta_0 + \boldsymbol{\beta}^T \boldsymbol{Y}_j]} \\ \Pr(K_j = 1 | \boldsymbol{Y}_j) = 1 - \Pr(K_j = 2 | \boldsymbol{Y}_j) &= \frac{1}{1 + \exp[\beta_0 + \boldsymbol{\beta}^T \boldsymbol{Y}_j]} \end{aligned} \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the set of logistic regression coefficients used for features. In combined notations the probability and the corresponding logged version have the form

$$\begin{aligned} \Pr(K_j = k | \boldsymbol{Y}_j) &= \frac{\exp[I_{j2}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{Y}_j)]}{1 + \exp[\beta_0 + \boldsymbol{\beta}^T \boldsymbol{Y}_j]} \\ \log[\Pr(K_j = k | \boldsymbol{Y}_j)] &= I_{j2}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{Y}_j) - \log[1 + \exp[\beta_0 + \boldsymbol{\beta}^T \boldsymbol{Y}_j]]. \end{aligned} \tag{4}$$

The penalized log likelihood function for n samples have the from:

$$\frac{1}{n}\sum_{j=1}^{n}\left[I_{j2}(\beta_0+\boldsymbol{\beta}^T\mathbf{Y}_j) - \log\left[1+\exp\left[\beta_0+\boldsymbol{\beta}^T\mathbf{Y}_j\right]\right]\right] + \lambda\left[[(1-\alpha)/2]||\boldsymbol{\beta}||_2^2+\alpha||\boldsymbol{\beta}||_1\right] \qquad (5)$$

where $||\boldsymbol{\beta}||_2^2 = \sum_{i=1}^{p}\beta_i^2$ and $||\boldsymbol{\beta}||_1 = \sum_{i=1}^{p}|\beta_i|$. The second term in the penalized likelihood represents the Elastic Net penalty. The penalized likelihood is maximized numerically for parameters (β_0, $\boldsymbol{\beta}$) for the given values of the penalty λ and split parameter α. The value $\alpha = 1$ corresponds to LASSO penalty introduced by Tibshirani [29] and $\alpha = 0$ corresponds to ridge regression penalty introduced by Hoerl and Kennard [28]. Due to the structure of the penalty outlined in (5) for all $\alpha > 0$ some coefficient estimates $\hat{\beta}_i$ are equal to zero, and the procedure serves as a variable selector [7]. The value of penalty λ is often estimated via a cross-validation procedure. This may be problematic in small sample sizes [37] [38] motivating us to examine the behavior across a range of values for the split parameter α. In this classification framework, the problem transforms into variable selection problem with p variables β_i and binary categorical outcome K_j. The Elastic Net approach directly accounts for the correlation among features.

Simulation studies

Simulation studies were performed to compare, Elastic Net, ridge regression, and LASSO to ANOVA, in the identification of features with respect to the following questions: 1) What is the effect of sample size (n)? 2) What is the effect of the correlation structure among features? 3) What is the effect of increasing the number of features (p)? 4) What is the impact of effect size (Δ)?

For the control ($k = 1$) and treatment ($k = 2$) groups twenty scenarios were considered in the simulation study (Table 1) covering small ($p = 205$) and larger ($p = 2050$) numbers of features with medium ($\rho = 0.4$) and high ($\rho = 0.8$) correlation between the causal features and non-causal features as well as independent features representing stochastic noise for a range of sample sizes from $n = 10$ to 100 where $n_1 = n_2$ and $n_1 + n_2 = n$.

All the samples were generated from the multivariate normal distribution (MVN) where the mean of each feature μ_i for $i = 1, 2, \ldots, p$ in the mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ was preliminary independently generated from the univariate gamma distribution. The gamma distribution had the shape parameter $\kappa = 50$, the scale parameter $\theta = 1/50$, and the density:

$$f(\mu_i|\kappa,\theta) = \frac{1}{\Gamma(\kappa)\theta^{\kappa}}\mu_i^{\kappa-1}e^{-\frac{\mu_i}{\theta}} \qquad (6)$$

where $\Gamma(\kappa)$ is the gamma function. In the given parametrization the mean of the gamma distributions for each μ_i is equal to $\kappa\theta = 1$ and variance is equal to $\kappa\theta^2 = 1/50$.

The variance-covariance matrix Σ used for the MVN simulation had a block diagonal structure with four independent blocks and diagonal variance values $\sigma^2 = 1$. The first three blocks of

Table 1. Summary of the simulation scenarios. For each scenario 1000 datasets were simulated. All together there were 20 scenarios considered.

Number of features (*p*)	Correlation (*ρ*)	Number of Samples (*n*)	Difference (Δ)
205	0.4	10	$\Delta_0 = 0.0$
2050	0.8	20	$\Delta_1 = 0.2$
		30	$\Delta_2 = 0.5$
		50	$\Delta_3 = 0.8$
		100	

https://doi.org/10.1371/journal.pone.0197910.t001

Σ consisted of 35 (or 350) features with the fixed correlation value ρ between the elements within each block. The fourth block had 100 (or 1000) features generated independently from the normal distribution to represent stochastic noise likely present in most omics experiments. The parameters of the distribution were $MVN(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ was the realization of the gamma distribution described in (6). Visualization of the correlation structure for $p = 205$ features is presented in Fig A in S1 Appendix. The structure for $p = 2050$ features is analogous. Twelve features were simulated with a difference in the means between the two treatment groups, four in each of the three correlated blocks with effect sizes $\Delta_1 = 0.2$, $\Delta_2 = 0.5$, and $\Delta_3 = 0.8$ for the first, the second, and the third block respectively. The effect size values $\Delta_1 = 0.2$, $\Delta_2 = 0.5$, and $\Delta_3 = 0.8$ according to Cohen [39] correspond to small, medium, and large effect sizes respectively. For each simulation scenario summarized in Table 1, 1000 datasets were generated and then analyzed with all methods. Simulation code is available in S1 File.

Analysis methods

In the ANOVA approach the difference between the treatment groups was considered significant for a given feature if the corresponding one-way ANOVA *F*-test resulted in a nominal *p*-value smaller than 0.05. The ANOVA approach was implemented in `R` using the function `aov` from the build-in core package `Stats`. The false discovery rate (FDR) method proposed by Benjamini and Hochberg [40] was used to account for multiple testing. The FDR adjustment was implemented using the `p.adjust` function also from core package `Stats` and the difference between the treatment groups was considered significant if the adjusted *p*-value was smaller than thresholds that were used: 0.05 and 0.20.

For the Elastic Net approach the logistic regression with Elastic Net penalty was fit using the `cv.glmnet()` function [33]. The default number of folds (10) was used for cross-validation for all scenarios. The features selected by the method were the ones that had non-zero coefficients β_i-s after the set was finalized. The Elastic Net method (5) depends on the choice of the penalty parameter α [7]. To investigate sensitivity of the results to values of α, analysis was performed on simulated data for α values 0 to 1, with 0.1 increment between the values. The α-sensitivity analysis was performed for each simulation scenario outlined in Table 1.

Real datasets used for methods illustration

The maize data consist of $n = 24$ samples with $n_1 = 12$ ambient and $n_2 = 12$ ozone samples. The goal was to identify metabolites *i*, that were different between the ambient and ozone treatment groups ($k = 1, 2$). Ozone is a phytotoxic air pollutant that enters plants through the stomatal pores on their leaves, and ultimately results in reduced crop yields [41]. The goal of the study was to identify specific metabolites and metabolic pathways impacted by the air pollutant. Metabolomics data were generated by the South East Center for Integrated Matabolomics (SECIM) and have been deposited to the Metabolomics Workbench [42] with project ID PR000193. After QC filtering using information from the blanks [43] there were 986 metabolites in positive ion mode and 863 metabolites in negative ion mode. After the model (1) was fit, the FDR correction was applied separately to the positive and negative ion mode data.

A gene expression dataset for type 1 diabetes was also considered, consisting of Illumina HiSeq 2000 RNA sequencing data for three lymphocyte cell types (CD19+ B cells, CD4+ T cells, CD8+ T cells) from 81 subjects with type 1 diabetes [44]. Gene expression was quantified as measurements of exon abundance as previously described, and only the subset of 163,713 exons detected in all three cell types was analyzed [44]. For the purposes of illustrating the approach presented in this manuscript, only expression data from CD4+ and CD8+ T cells from the 79 individuals with complete data were analyzed to identify the set of differentially-expressed exons between CD4+ and CD8+ T cells in type 1 diabetes cases. Only those exons with average depths per nucleotide (APN) $\geq$5 were used for feature selection resulting in 8,268 exons that were present in all samples. Data are deposited in dbGaP [45] under accession number phs001426.v1.p1.

Results

Simulated data

For the Elastic Net approach simulated data were evaluated across the range of possible values for the split parameter (Table 1). The parameter α that is responsible for a penalty split is referred to as a higher level parameter in the package documentation [36]. This α-sensitivity analysis was used to determine the values of α for use in comparison to the ANOVA approach. A subset of representative results is presented in Fig 1 ($\rho = 0.4$) and Fig 2 ($\rho = 0.8$). The complete set of results is provided in S1 Appendix Figs B–G. The summary statistics for the simulated data for sample sizes 10 and 100 are summarized in S1 Appendix Tables A–H. When fitting the Elastic Net model (5) there is also an option to cross validate not only the penalty value λ but also the penalty split parameter α [7]. When performing such cross validation the estimated value of α will be different for every dataset, which makes interpretation and justification of the specific penalty split parameter choice challenging. The possibility to over fit for small sample sizes should also be a concern [37] [38]. In this work a range of values for α was examined with the goal of selecting a single value.

For all simulation scenarios and for all values of α except 0 an increase in the effect size increased the power of detection. The increase in the sample size also increased the power of detection for all effect sizes and values of α. As the sample size increases the gain in power for small effect sizes is modest compared to large effect sizes for all methods. Power of detection decreased monotonically with increase in the penalty split value α from ridge regression ($\alpha =$ 0) to LASSO ($\alpha = 1$). Ridge regression ($\alpha = 0$) included all variables and resulted in estimates of power and Type I error of 1. LASSO ($\alpha = 1$) had the lowest power of all values of α for each scenario and was in the range 0.7 − 0.8 only for the larger sample and effect sizes ($\Delta = 0.8$ and $n = 100$).

For ($\alpha > 0$) an increase in the sample size increased the Type I error. The increase was more pronounced for the smaller number of features ($p = 205$) than for the larger number of features ($p = 2050$). Type I error decreased monotonically with increase in the penalty split value α from ridge regression to LASSO. However, this is not a linear function. For ridge penalty ($\alpha = 0$) coefficients of the correlated predictors are shrunk towards each other, for LASSO penalty ($\alpha = 1$) only single coefficient of the correlated group selects one and discards the others [7]. The value $\alpha = 0.5$ that was used for the analysis in this work tends to either select or not select the groups of correlated features together [36]. For the larger number of features ($p = 2050$) the steep reduction in Type I error occurs at much lower values of α than for the smaller number of features. The larger value of correlation values ($\rho = 0.8$ vs $\rho = 0.4$) increases Type I errors for larger sample sizes. For any value of α, Type I error was larger than nominal level 0.05 when the number of features was small ($p = 205$) and the sample size is $n = 100$. For the larger number of features ($p = 2050$) and small sample sizes ($n = 10, 20, 30, 50$) the Type I error was close to nominal level for $\alpha \in [0.5; 1]$ including for the LASSO. Based on these results $\alpha = 0.5$ was considered a balanced representation of power and Type I error for the Elastic Net approach and was used for comparison with the ANOVA approach.

Type I error for the Elastic Net $\alpha = 0.5$ and LASSO increased monotonically with the increase in the sample size but still stayed within the nominal level (0.05) for the considered sample sizes ($n \leq 100$) if the number of features was larger ($p = 2050$). For the smaller number

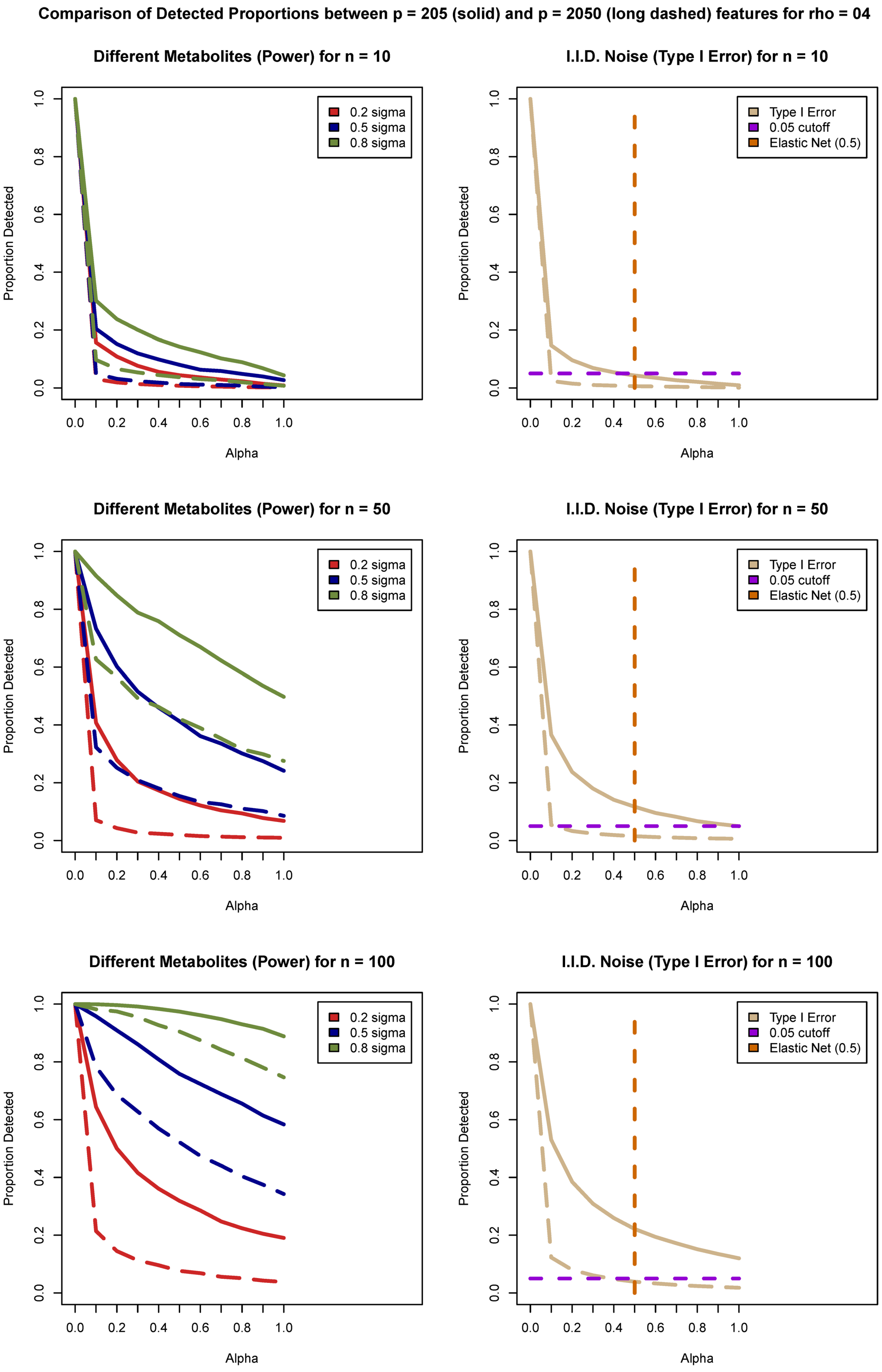

Fig 2. Visualization of power (left column) and Type I error (right column) estimates. Comparison between $p = 205$ (solid lines) and $p = 2050$ (dashed line) features for $\rho = 0.8$ and sample sizes $n = 10$ (top row), $n = 50$ (middle row), and $n = 100$ (bottom row). The value of the penalty split parameter α is plotted on the x-axis. Type I error and power estimates are plotted on y-axis for the values of α in the range of [0; 1] with 0.1 increments. In the left column power estimates are provided based on the four different features for each of the effect sizes ($\Delta_1 = 0.2$ is the red line, $\Delta_2 = 0.5$ is the blue line, and $\Delta_3 = 0.8$ is the green line). In the right column Type I error estimates are provided (beige lines) based

on the random noise features together with a 0.05 threshold plotted as a purple dashed line. The vertical dashed line in the right column plots corresponds to penalty split value $\alpha = 0.5$. The value of $\alpha = 0$ corresponds to ridge regression and $\alpha = 1$ corresponds to LASSO.

https://doi.org/10.1371/journal.pone.0197910.g002

of features ($p = 205$) and sample sizes above fifty the error was above nominal level (0.05). To further investigate the behavior of Type I error based on the sample size for the larger number of features, the sizes $n = 200$, 500 and 1000 were considered for $p = 2050$ features and correlation $\rho = 0.4$. The results indicated that the Type I error kept increasing monotonically with the sample size increase for both LASSO and Elastic Net and were above the nominal level of 0.05 for sample size $n = 200$ for Elastic Net and for sample size $n = 500$ for LASSO.

The ANOVA tests were performed independently on each feature, the Type I error and the power of the tests were very similar regardless of the number of features and correlation between features. As expected, as sample size and effect size increased power increased. The power of ANOVA approach was in the range: $0.7 - 0.8$ for sample sizes greater than $n = 50$ and effect size $\Delta = 0.8$ *regardless* of the number of features. The results for Type I error were at the nominal level (0.05) for unadjusted ANOVA despite the multiple testing and as expected [40] were well below the nominal level of 0.2 after FDR adjustment. Indeed, after FDR adjustment at 0.20 the results were below 0.05.

For features simulated as correlated with differentially expressed features, but not themselves simulated to have a difference in the mean, the Elastic Net and LASSO approaches over selected features when the number of features was $p = 205$ and sample size was $n \geq 50$. As expected, the performance of the ANOVA for these features mirrored the behavior of the Type I error for all sample sizes. Power estimates for sample sizes smaller than 50 were lower than 0.5 for small effect sizes for all tests. The increase in the effect size Δ increases power of the tests. Overall, LASSO performed worse than either Elastic Net or ANOVA (adjusted or unadjusted).

Of particular interest, the increase in the number of features from $p = 205$ to $p = 2050$ affected the Type I error and the power of any of the considered methods much less than the increase in number of samples from $n = 10$ to $n = 100$. This indicates that the number of features examined does not play as crucial a role as the correlation and dependency structure in the data and the sample size. The results for $\rho = 0.4$ are provided in Figs 3 and 4 and for $\rho = 0.8$ in S1 File Figs F and G.

Real data

Venn diagrams (Fig 5, S1 Appendix Fig H) of the result comparison for different methods for maize data with FDR adjustment for ANOVA are presented in Fig 5 and without FDR adjustment in S1 Appendix Fig H. The results for the positive and negative ion modes were similar. When Type I error is minimized (FDR level 0.05) ANOVA selects comparable number of features to the Elastic Net and LASSO, but the three methods do not have perfect overlap between the set of features and none of the sets include the other. At an FDR level 0.2 the set of features selected by ANOVA does not include all the features selected by LASSO and Elastic Net. This results hold for both positive and negative ion modes. Interestingly, in the negative ion mode there were features selected by LASSO ($p = 3$) and not selected by Elastic Net. For positive ion model features selected by LASSO formed a proper subset of the features selected by Elastic Net.

In the diabetes data (Figs I and J in S1 Appendix) a set of $p = 1781$ exons was selected using the ANOVA method and FDR level 0.2 while the Elastic Net procedure selected a subset of 58

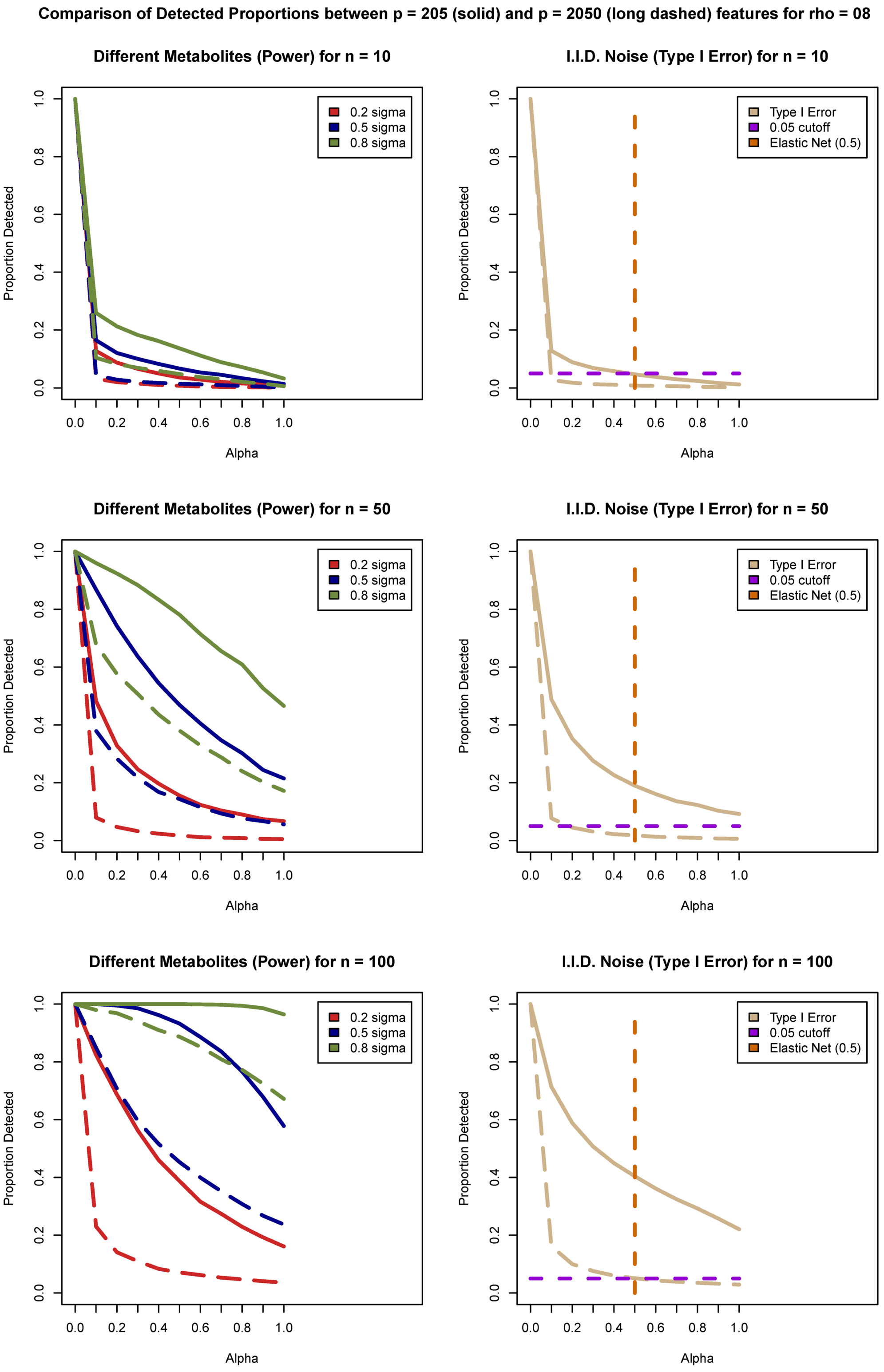

Fig 3. Visualization of power and Type I error estimates comparison for p = 205 features, correlation ρ = 0.4, and all sample sizes. Each row of the plots corresponds to a feature selection method. ANOVA FDR adjustment cutoff is 0.2. The value of the sample size (n) is displayed on the x-axis in all plots. The estimates of power and Type I error are provided on the y-axis. In the left column power estimates are provided based on the four different features for each of the effect sizes ($\Delta_1 = 0.2$ is the red line, $\Delta_2 = 0.5$ is the blue line, and $\Delta_3 = 0.8$ is the green line). In the right column Type I error estimates are provided (beige lines) based on the random noise features together with a 0.05 threshold plotted as

a purple dashed line. In the middle column the proportions of non-different detected features within each block correlated to different ones for each of the blocks and corresponding effect sizes ($\Delta_1 = 0.2$ is the red line, $\Delta_2 = 0.5$ is the blue line, and $\Delta_3 = 0.8$ is the green line) are displayed.

https://doi.org/10.1371/journal.pone.0197910.g003

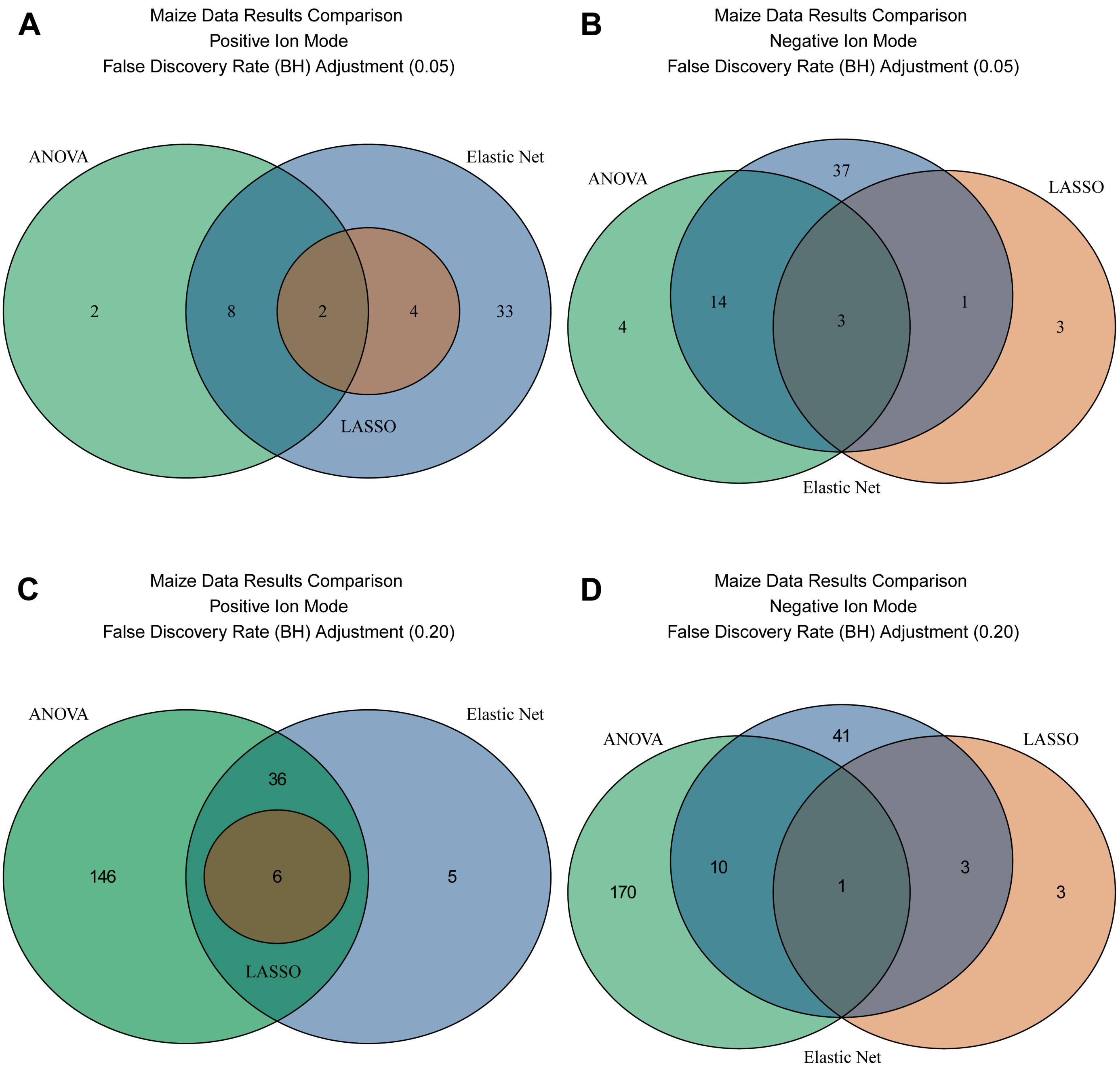
A
Maize Data Results Comparison
Positive Ion Mode
False Discovery Rate (BH) Adjustment (0.05)
ANOVA
Elastic Net
2
8
2
4
33
LASSO
B
Maize Data Results Comparison
Negative Ion Mode
False Discovery Rate (BH) Adjustment (0.05)
ANOVA
37
LASSO
4
14
3
1
3
Elastic Net
C
Maize Data Results Comparison
Positive Ion Mode
False Discovery Rate (BH) Adjustment (0.20)
ANOVA
Elastic Net
36
146
6
5
LASSO
D
Maize Data Results Comparison
Negative Ion Mode
False Discovery Rate (BH) Adjustment (0.20)
ANOVA
41
LASSO
170
10
1
3
3
Elastic Net

Fig 4. Visualization of power and Type I error estimates comparison for $p = 2050$ features, correlation $\rho = 0.4$, and all sample sizes. Each row of the plots corresponds to a feature selection method. ANOVA FDR adjustment cutoff is 0.2. The value of the sample size (n) is displayed on the x-axis in all plots. The estimates of power and Type I error are provided on the y-axis. In the left column power estimates are provided based on the four different features for each of the effect sizes ($\Delta_1 = 0.2$ is the red line, $\Delta_2 = 0.5$ is the blue line, and $\Delta_3 = 0.8$ is the green line). In the right column Type I error estimates are provided (beige lines) based on the random noise features together with a 0.05 threshold plotted as a purple dashed line. In the middle column the proportions of non-different detected features within each block correlated to different ones for each of the blocks and corresponding effect sizes ($\Delta_1 = 0.2$ is the red line, $\Delta_2 = 0.5$ is the blue line, and $\Delta_3 = 0.8$ is the green line) are displayed.

https://doi.org/10.1371/journal.pone.0197910.g004

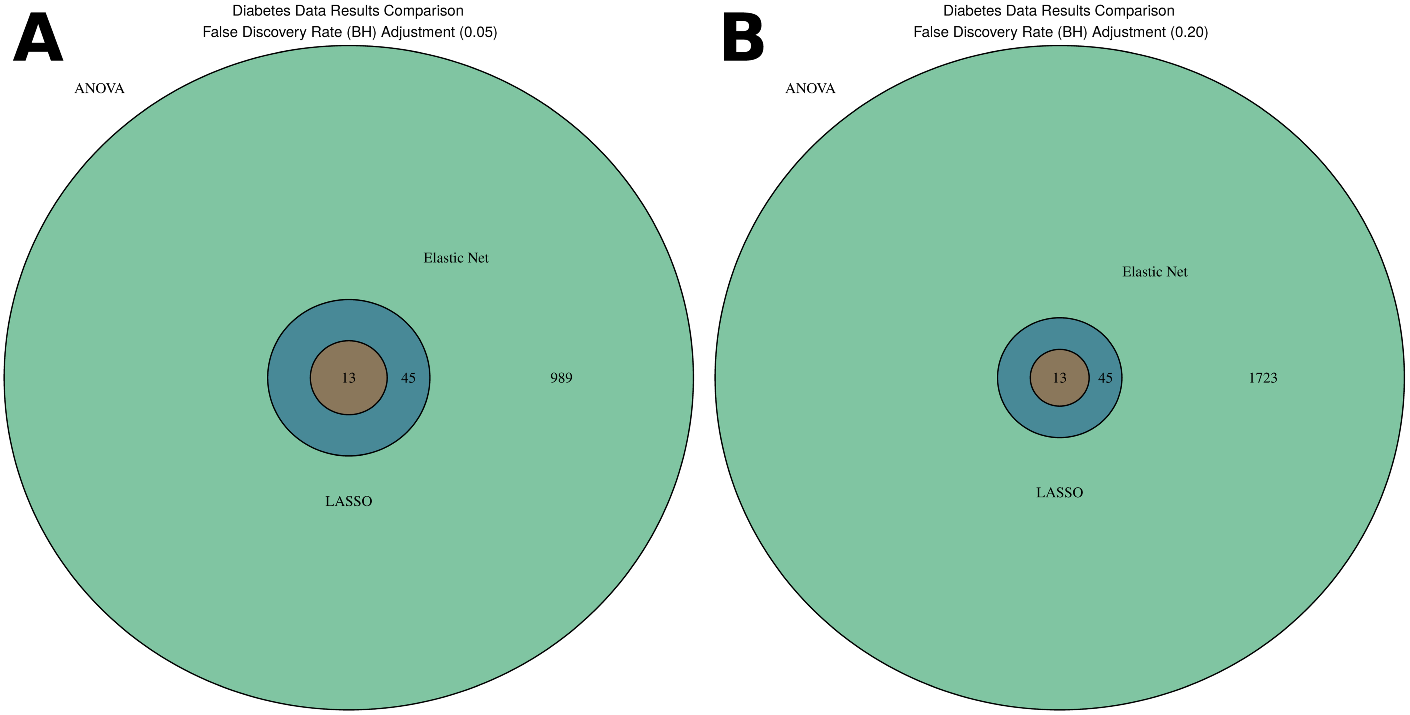

Fig 5. Venn diagrams depicting the results for the maize data. ANOVA (Green), Elastic Net (Blue) and LASSO (Brown) are compared. In Panel A the positive ion mode is shown with an FDR for the ANOVA of 0.05 while in Panel B the negative ion mode is show. Panel C is the positive ion mode with FDR of 0.2 and Panel D is the negative ion mode.

https://doi.org/10.1371/journal.pone.0197910.g005

exons and the LASSO procedure selected 13 exons that formed a subset of Elastic Net selection. For the ANOVA method and FDR level 0.05 the set of p = 1047 exons were selected, and, again, the exons selected by LASSO formed a subset of exons selected by Elastic Net, which was in its turn the subset of the ANOVA features. The unadjusted ANOVA results are provided in S1 Appendix Fig I.

Discussion

For omics data the identification of features that are potentially different among groups is a common analytical goal. Typically biologists view this initial experiment as a screening step although this is often not explicitly stated. It is not a realistic analytical goal to identify a biomarker in a single study, particularly a single study with sample sizes that are small. Many omics experiments have sample sizes in the tens. Further, there are many levels of structure in these data. There are biological correlations among the features examined due to the shared pathways that generate the observations and the experiments themselves often have a structure due to batch effects during data acquisition and potential covariates. For human studies, the initial sample collections may be complicated by patient characteristics such as age, sex or comorbidity. These analytic concerns should be considered in the evaluation of statistical methods for variable selection [46] [47] [48]. Here we explicitly consider the goal of the experiment to be the identification of a reasonable number of compounds that can be followed up. That is an explicit goal to minimize Type II error without undue inflation of the Type I error.

Methods have been developed for variable selection in omics data with the stated goal of finding the biomarker [6] [49] [50] [14]. However, it has been established that popular partial least squares discriminant analysis (PLS-DA) based approaches often overfit [51] [52].

Similarly machine learning approaches [53] [54] have been demonstrated to overfit for small samples. Elastic Net has been consistently found to perform well compared to other machine learning approaches [7] [6].

Elastic Net requires the selection of the penalty split α which can either be specified [36] [55] or, potentially, cross validated. There are several issues with cross validation in this context. First, cross validation for sample sizes less than 100 may be problematic due to poor performance [37] [38]. In this context cross-validation across the values of α in addition to the cross-validation over the values of λ is likely to perform poorly. Further, if cross-validation is performed for the penalty split parameter α, the analysis for each dataset will have a different value for $\hat{\alpha}$ which makes the direct comparison of the model challenging.

For small sample sizes when the number of features is also small, such as in a targeted panel of metabolites, the Elastic Net has an inflated Type I error. For larger numbers of features ($n < p$) the Elastic Net at small sample sizes (<100) has Type I error control, but has lower power than the uncorrected ANOVA. The uncorrected ANOVA does not have an inflation of the Type I error in these scenarios. Type I error inflation (overselection) for sample size of 50 has been reported for LASSO and other penalized approaches [56]. Ridge regression, will include all features in the prediction and has a Type I and Type II error of 1. The Elastic Net has also been shown to have a higher than nominal Type I error in studies with fairly large samples by omics standards (n = 500) [57].

Elastic Net has other distinct drawbacks in comparison to ANOVA as an initial screening tool. Elastic Net lacks an analytical solution [30] [31] making interpretation of the coefficient values based on values of the original measurement challenging. Further, some of the elements of the experiment may not be accounted for properly in normalization [58] [59] [60]. This may result in Elastic Net picking features that are (very) different between unaccounted for sources of variation (e.g. batch) and ignoring important but potentially smaller difference between treatments. This may explain why in the maize data there are features selected by Elastic Net that are not selected by the ANOVA approach. There is a possibility to include batch ID and other relevant variables as features that are subject to selection, but this is not expected to have the same effect as including covariates in an ANOVA model.

Often initial untargeted metabolomics studies are used to cast a broad net in identifying what set of compounds may be different among groups [61]. This is analogous to a screening study where the cost of the false negative should be weighed higher than the cost of the false positive, as subsequent experiments will remove the false positives. There is always a tradeoff [62] [63] [64]. As with screening studies, the follow up studies used to confirm a difference between groups are more expensive in labor, time and materials than the untargeted global approaches indicating that Type I error cannot be completely disregarded. The identification of groups of features that are differentially expressed from a shared pathway will provide insight into underlying biological processes [65] [66] [67]. Ideally correlated features that are truly differentially expressed should be selected and correlated features that are not different should not be selected. The ANOVA approach does not have an elevated Type I error rate among features correlated to those that are differentially expressed, compared to independent random noise, a desirable property.

The Type II error for the ANOVA approach was, in some circumstances, lower than the Elastic Net and LASSO. The conditions under which a two sample t-test can be used to extract all of the features that are different between conditions has recently been examined [68]. Fan and Fan prove that all features different between the two groups are selected with a probability of one (that is the probability of Type II errors goes to zero) assuming that most of the features are not different between the groups, Cramer's condition holds and the variance for each feature is bounded away from zero. The diabetes data here clearly violate the first condition, as the number of genes different between the two conditions is large. However, in this case the ANOVA approach selected more features than the Elastic Net and the LASSO suggesting that the Type II error when the assumption of sparse differences is violated may not affect the Type II error.

The ultimate goal of a biomarker identification is to have one or a few biomarkers capable of predicting the treatment group. However, minimizing features at an early stage, using data from a single, small experiment risks the exclusion of the optimal biomarker. For biomarker identification to be successful, it should be thought of as a sequential screening procedure, a result of a preponderance of evidence over time. In the initial stages, selection of all of the significant correlated features (low Type II error), and none of the non-significant correlated features (low Type I error) allows the scientist to examine sets of significant features for biological insights. Enrichment analyses and other approaches can lend insights into the biological pathways and can help design subsequent experiments to measure those pathways more completely- for example developing a targeted panel of metabolites to be assayed in a large population based on the indication that the biological pathway is affected by the disease.

Conclusion

An ANOVA is a simple technique, can account for a complex set of experimental design conditions [59], is likely to achieve a very low Type II error [68] and in the conditions studied here has lower Type I error than other more complex methods. For studies whose goal is to advance a set of features to the next round of testing for biological relevance, an ANOVA is an excellent choice.

Supporting information

S1 Appendix. The supplementary material contains tables and figures of the result summaries for the simulated and real data analysis that were not included in the main text due to the manuscript size limitations.
(PDF)

S1 File. The code for the simulations.
(R)

Acknowledgments

This work was funded by multiple sources that authors want to acknowledge: National Institutes of Health (NIH) U24 DK097209, National Science Foundation (NSF) 1238030, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) DP3DK085678, and the University of Florida Informatics Institute (UFII) Fellowship.

Author Contributions

Conceptualization: Lauren M. McIntyre.

Data curation: Jeremy R. B. Newman, Lauren M. McIntyre.

Formal analysis: Alexander Kirpich.

Funding acquisition: Elizabeth A. Ainsworth, George Michailidis, Lauren M. McIntyre.

Project administration: Lauren M. McIntyre.

Resources: Elizabeth A. Ainsworth, Jessica M. Wedow, Jeremy R. B. Newman, Lauren M. McIntyre.

Supervision: Lauren M. McIntyre.

Writing – original draft: Alexander Kirpich, Lauren M. McIntyre.

Writing – review & editing: Alexander Kirpich, Elizabeth A. Ainsworth, Jessica M. Wedow, Jeremy R. B. Newman, George Michailidis, Lauren M. McIntyre.

References

1. Katajamaa M, Miettinen J, Orešič M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics. 2006; 22(5):634–636. https://doi.org/10.1093/bioinformatics/btk039 PMID: 16403790
2. Dunn WB, Wilson ID, Nicholls AW, Broadhurst D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. 2012;.
3. Dunn WB, Erban A, Weber RJ, Creek DJ, Brown M, Breitling R, et al. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics. 2013; 9(1):44–66. https://doi.org/10.1007/s11306-012-0434-4
4. Creek DJ, Dunn WB, Fiehn O, Griffin JL, Hall RD, Lei Z, et al. Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics. 2014; 10(3):350. https://doi.org/10.1007/s11306-014-0656-8
5. Alonso A, Marsal S, Julià A. Analytical methods in untargeted metabolomics: state of the art in 2015. Frontiers in bioengineering and biotechnology. 2015; 3:23. https://doi.org/10.3389/fbioe.2015.00023 PMID: 25798438
6. Acharjee A, Finkers R, Visser RG, Maliepaard C. Comparison of regularized regression methods for omics data. Metabolomics. 2013; 3(126):2153–0769.
7. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
8. Molins CR, Ashton LV, Wormser GP, Hess AM, Delorey MJ, Mahapatra S, et al. Development of a metabolic biosignature for detection of early Lyme disease. Clinical Infectious Diseases. 2015; p. civ185. https://doi.org/10.1093/cid/civ185 PMID: 25761869
9. Cho S, Kim H, Oh S, Kim K, Park T. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. In: BMC proceedings. vol. 3. BioMed Central; 2009. p. S25. https://doi.org/10.1186/1753-6561-3-s7-s25
10. Koulman A, Tapper BA, Fraser K, Cao M, Lane GA, Rasmussen S. High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics. Rapid Communications in Mass Spectrometry. 2007; 21(3):421–428. https://doi.org/10.1002/rcm.2854 PMID: 17206744
11. Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC bioinformatics. 2011; 12(1):253. https://doi.org/10.1186/1471-2105-12-253 PMID: 21693065
12. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences. 2002; 99(10):6567–6572. https://doi.org/10.1073/pnas.082099299
13. Grissa D, Pétéra M, Brandolini M, Napoli A, Comte B, Pujos-Guillot E. Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. Frontiers in molecular biosciences. 2016; 3. https://doi.org/10.3389/fmolb.2016.00030 PMID: 27458587
14. Ju H, Brasier AR. Variable selection methods for developing a biomarker panel for prediction of dengue hemorrhagic fever. BMC research notes. 2013; 6(1):365. https://doi.org/10.1186/1756-0500-6-365 PMID: 24025735
15. Li F, Shin R, Paxson V. Exploring privacy preservation in outsourced k-nearest neighbors with multiple data owners. In: Proceedings of the 2015 ACM Workshop on Cloud Computing Security Workshop. ACM; 2015. p. 53–64.
16. Meyer O, Bischl B, Weihs C. Support vector machines on large data sets: Simple parallel approaches. In: Data Analysis, Machine Learning and Knowledge Discovery. Springer; 2014. p. 87–95.
17. Ripley B, Venables W, Ripley MB. Package "nnet". R package version. 2016; p. 7–3.

18. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv preprint arXiv:160600930. 2016;.

19. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Briefings in bioinformatics. 2017;. https://doi.org/10.1093/bib/bbx124 PMID: 29045534

20. Kursa MB, Rudnicki WR, et al. Feature selection with the Boruta package. J Stat Softw. 2010; 36(11):1–13. https://doi.org/10.18637/jss.v036.i11

21. Janitza S, Celik E, Boulesteix AL. A computationally fast variable importance test for random forests for high-dimensional data. Advances in Data Analysis and Classification. 2015; p. 1–31.

22. Szymczak S, Holzinger E, Dasgupta A, Malley JD, Molloy AM, Mills JL, et al. r2VIM: A new variable selection method for random forests in genome-wide association studies. BioData mining. 2016; 9(1):7. https://doi.org/10.1186/s13040-016-0087-3 PMID: 26839594

23. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010; 26(10):1340–1347. https://doi.org/10.1093/bioinformatics/btq134 PMID: 20385727

24. Lu F, Petkova E. A comparative study of variable selection methods in the context of developing psychiatric screening instruments. Statistics in medicine. 2014; 33(3):401–421. https://doi.org/10.1002/sim.5937 PMID: 23934941

25. Bellino A, Colombo C, Iovieno P, Alfani A, Palumbo G, Baldantoni D. Chemometric technique performances in predicting forest soil chemical and biological properties from UV-Vis-NIR reflectance spectra with small, high dimensional datasets. iForest-Biogeosciences and Forestry. 2015; 9(1):101. https://doi.org/10.3832/ifor1495-008

26. Bujak R, Daghir-Wojtkowiak E, Kaliszan R, Markuszewski MJ. PLS-based and regularization-based methods for the selection of relevant variables in non-targeted metabolomics data. Frontiers in molecular biosciences. 2016; 3. https://doi.org/10.3389/fmolb.2016.00035 PMID: 27508208

27. Morozova O, Levina O, Uusküla A, Heimer R. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC medical research methodology. 2015; 15(1):71. https://doi.org/10.1186/s12874-015-0066-2 PMID: 26319135

28. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. Technometrics. 1970; 12(1):69–82. https://doi.org/10.1080/00401706.1970.10488634

29. Tibshirani R. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society Series B (Methodological). 1996; p. 267–288.

30. Efron B, Hastie T, Johnstone I, Tibshirani R, et al. Least angle regression. The Annals of statistics. 2004; 32(2):407–499. https://doi.org/10.1214/009053604000000067

31. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010; 33(1):1. https://doi.org/10.18637/jss.v033.i01 PMID: 20808728

32. Zhou Q, Chen W, Song S, Gardner JR, Weinberger KQ, Chen Y. A Reduction of the Elastic Net to Support Vector Machines with an Application to GPU Computing. In: AAAI; 2015. p. 3210–3216.

33. R Package `glmnet`. http://cranr-projectorg/web/packages/glmnet/glmnetpdf Accessed May 25, 2017. 2017;.

34. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science. 1999; 286 (5439):531–537. https://doi.org/10.1126/science.286.5439.531 PMID: 10521349

35. Neto EC, Bare JC, Margolin AA. Simulation studies as designed experiments: the comparison of penalized regression models in the”large p, small n” setting. PloS one. 2014; 9(10):e107957. https://doi.org/10.1371/journal.pone.0107957

36. Hastie T, Qian J. Glmnet Vignette; 2014.

37. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004; 20(3):374–380. https://doi.org/10.1093/bioinformatics/btg419 PMID: 14960464

38. Isaksson A, Wallman M, Göransson H, Gustafsson MG. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters. 2008; 29(14):1960–1965. https://doi.org/10.1016/j.patrec.2008.06.018

39. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd Edition). 1988; p. 20–27.

40. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; p. 289–300.

41. McGrath JM, Betzelberger AM, Wang S, Shook E, Zhu XG, Long SP, et al. An analysis of ozone damage to historical maize and soybean yields in the United States. Proceedings of the National Academy of Sciences. 2015; 112(46):14390–14395. https://doi.org/10.1073/pnas.1509777112

42. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic acids research. 2015; p. gkv1042. https://doi.org/10.1093/nar/gkv1042 PMID: 26467476

43. Patterson RE, Kirpich AS, Koelmel JP, Kalavalapalli S, Morse AM, Cusi K, et al. Improved Experimental Data Processing for UHPLC-HRMS/MS Lipidomics Applied to Fatty Liver Disease. Metabolomics. 2017;. https://doi.org/10.1007/s11306-017-1280-1

44. Newman JRB, Conesa A, Mika M, New FN, Onengut-Gumuscu S, Atkinson MA, et al. Disease-specific biases in alternative splicing and tissue-specific dysregulation revealed by multi-tissue profiling of lymphocyte gene expression in type 1 diabetes. Genome Research Accepted pending revisions. 2017;. https://doi.org/10.1101/gr.217984.116

45. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nature genetics. 2007; 39(10):1181–1186. https://doi.org/10.1038/ng1007-1181 PMID: 17898773

46. Enot DP, Lin W, Beckmann M, Parker D, Overy DP, Draper J. Preprocessing, classification modeling and feature selection using flow injection electrospray mass spectrometry metabolite fingerprint data. Nature Protocols. 2008; 3(3):446. https://doi.org/10.1038/nprot.2007.511 PMID: 18323816

47. De Maturana EL, Ibáñez-Escriche N, González-Recio Ó, Marenne G, Mehrban H, Chanock SJ, et al. Next generation modeling in GWAS: comparing different genetic architectures. Human genetics. 2014; 133(10):1235–1253. https://doi.org/10.1007/s00439-014-1461-1

48. Maturana E, Pineda S, Brand A, Steen K, Malats N. Toward the integration of Omics data in epidemiological studies: still a"long and winding road". Genetic epidemiology. 2016; 40(7):558–569. https://doi.org/10.1002/gepi.21992

49. Ghosh D, Chinnaiyan AM. Classification and selection of biomarkers in genomic data using LASSO. BioMed Research International. 2005; 2005(2):147–154.

50. McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Hafen R, Ramey J, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert opinion on medical diagnostics. 2013; 7(1):37–51. https://doi.org/10.1517/17530059.2012.718329 PMID: 23335946

51. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. Journal of Chemometrics. 2014; 28(4):213–225. https://doi.org/10.1002/cem.2609

52. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. Analytica chimica acta. 2015; 879:10–23. https://doi.org/10.1016/j.aca.2015.02.012 PMID: 26002472

53. Han H, Jiang X. Overcome support vector machine diagnosis overfitting. Cancer informatics. 2014; 13(Suppl 1):145. https://doi.org/10.4137/CIN.S13875 PMID: 25574125

54. Segal MR. Machine learning benchmarks and random forest regression. Center for Bioinformatics & Molecular Biostatistics. 2004;.

55. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, Inference and Prediction: Second Edition. New York: Springer. 2009; p. 661–666.

56. Kwon S, Oh S, Lee Y. The use of random-effect models for high-dimensional variable selection problems. Computational Statistics & Data Analysis. 2016; 103:401–412. https://doi.org/10.1016/j.csda.2016.05.016

57. Vasquez MM, Hu C, Roe DJ, Chen Z, Halonen M, Guerra S. Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. BMC medical research methodology. 2016; 16(1):154. https://doi.org/10.1186/s12874-016-0254-8 PMID: 27842498

58. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nature genetics. 2001; 29 (4):389. https://doi.org/10.1038/ng766 PMID: 11726925

59. Chu TM, Weir B, Wolfinger R. A systematic statistical linear modeling approach to oligonucleotide array experiments. Mathematical biosciences. 2002; 176(1):35–51. https://doi.org/10.1016/S0025-5564(01)00107-9 PMID: 11867082

60. Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YLS, Sultana H, et al. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. BMC genomics. 2016; 17(1):28. https://doi.org/10.1186/s12864-015-2353-z PMID: 26732976

61. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nature reviews Molecular cell biology. 2012; 13(4):263–269. https://doi.org/10.1038/nrm3314 PMID: 22436749

62. Fisher RA. Statistical methods for research workers. Genesis Publishing Pvt Ltd; 1925.

63. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of eugenics. 1936; 7(2):179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

64. Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics. 2013; 9(2):280–299. https://doi.org/10.1007/s11306-012-0482-9 PMID: 23543913

65. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2009; 4(1):44–57. https://doi.org/10.1038/nprot.2008.211

66. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009; 25(2):288–289. https://doi.org/10.1093/bioinformatics/btn615 PMID: 19033274

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

68. Fan J, Fan Y. High dimensional classification using features annealed independence rules. Annals of statistics. 2008; 36(6):2605. https://doi.org/10.1214/07-AOS504 PMID: 19169416