Prediction of CYP3A4 metabolic activity from whole genome RNA-seq data with feature selection machine learning methods

Yichen Jia

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2017

Reading Committee:

Timothy A. Thornton, PhD; Chair

Michael C. Wu, PhD

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

Prediction of CYP3A4 metabolic activity from whole genome RNA-seq data with feature selection machine learning methods

Yichen Jia

Chair of the Supervisory Committee:
Timothy A. Thornton, Robert W. Day Endowed Professor of Public Health, Associate Professor
Department of Biostatistics

CYP3A4, one of the isozyme of the cytochromes P450 (CYPs), contributes significantly to drug clearance and drug-drug interactions. The goals of this project are to identify hepatically-expressed genes that are associated with CYP3A4 metabolic activity in human liver tissue and to predict CYP3A4 activity using gene expression data from whole genome RNA sequences. Due to the high-dimensionality of the dataset, we applied lasso and elastic net, two feature selection machine learning methods, for prediction and graphical lasso was used for constructing gene network graphs. A simulation study was performed to assess the performance of the prediction algorithms and to evaluate the efficiency of gene selection using the machine learning methods. We assessed prediction performance based on correlations, and the correlation between measured CYP3A4 activity and predicted activity was approximately 0.4 and 0.5 when reductase

was excluded and included, respectively, for both lasso and elastic net. In addition to the CYP3A4 gene, we also identified the GZMA gene as a strong candidate for prediction of CYP3A4 activity that should be investigated in future studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to my advisor, Dr. Timothy A. Thornton, for his support and guidance on this thesis. Dr. Michael C. Wu, who also served on the reading committee, has similarly given suggestion that have been helpful for this project. I also thank Dr. Kenneth E. Thummel and Dr. Katrina G. Claw from the UW Department of Pharmaceutics and Dr. Erin Schuetz and Dr. Amarjit Chaudhry at St. Jude Children's Research Hospital for providing and preparing the dataset I used in this thesis. Finally, I am thankful for my family who have been unconditionally supporting me throughout my education.

Chapter 1. INTRODUCTION

The cytochromes P450 (CYPs) constitute a major drug metabolizing enzyme family that catalyzes the metabolism of many medicines and endogenous compounds (Zanger and Schwab, 2013). Of the 57 human CYP isozymes, five isozymes - 1A2, 2C9, 2C19, 2D6, and 3A4 metabolize approximately 80% known drugs in humans. Of these, CYP3A4 stands out in terms of the extent of its contribution to drug clearance and its involvement in clinically significant drug-drug interactions (Ortiz de Montellano, 2005). There is also considerable inter-individual variability in constitutive CYP3A4 activity that further confounds safe and efficacious treatment with CYP3A4 substrates (Ozdemir et al., 2000). It has previously been reported that genetic factors contribute to variability in CYP3A4 activity (Lamba et al. 2010); however, very little of this variability can actually be attributed to known genetic variation (SNVs or CNVs) in the CYP3A4 gene itself. The goals of this project are to identify hepatically-expressed genes that are associated with the CYP3A4 metabolic activity in human liver tissue and to predict CYP3A4 activity with gene expression data from whole genome RNA sequences. The overarching motivation for this project is to provide new insight into possible pathways involved in CYP3A4 activity regulation and to expand the identification of transcriptome variation contributing to inter-individual differences in CYP3A4 activity for future studies.

This project involves a sample of 231 human liver tissues with whole genome RNA sequences, expression levels for more than 20,000 genes and microsomal CYP3A4 activity. Demographic and clinical data are also available for each sample. A significant challenge in the analysis of this data is that the number of genes with expression measurements is substantially

larger than the number of liver samples. As a result, it is expected that this high-dimensional dataset will contain many genes whose expression levels are not truly associated with CYP3A4 activity, which increases the risk of overfitting the data as well as identifying genes that are false positives. As a result, there are significant challenges in developing an effective prediction model for CYP3A4 activity from whole genome RNA-seq data. In addition, for high-dimensional data, multicollinearity among the predictors can lead to highly variable regression coefficients. Thus, machine learning algorithms with feature selection are essential for obtaining a parsimonious prediction model that provides useful information about the relationship between CYP3A4 activity and the high-dimensional gene expression values, as well as for enhancing the prediction performance.

Various regularization methods have recently been developed to deal with the challenges mentioned above for the analysis of high-dimensional data, including ridge regression (Hoerl and Kennard, 1970), lasso regression (Tibshirani, 1996), and the elastic net (Zou and Hastie, 2005), which are all regression-based methods. These popular methods control model complexity by imposing regularity and shrinking regression coefficients of predictors towards zero. In this study, we applied both lasso and elastic net algorithm for feature selection. Ridge regression was not performed because the final model would contain all variables, which is not useful for excluding genes from the model that are not involved with the outcome of interest. Lasso regression, on the other hand, shrinks most coefficients to 0, resulting in a sparse model which is of interest for our study. However, lasso has some limitations. First, if a group of predictors are highly correlated, lasso tends to select only one of them and ignore the others. Second, lasso cannot select more variables than observations, which could be a problem in studies for genomic prediction. To overcome the drawback of ridge and lasso, elastic net linearly combines the

penalties of the ridge and lasso regression which facilitates selection of groups of correlated features and enables the selection of more features than samples.

In this study, we also construct gene networks using genes that have significant marginal effects on CYP3A4 metabolic activity. A gene network can be viewed as an undirected graph where each vertex of the graph represents a gene and the edge represents the association relation between a pair of genes. In our analyses, gene expression values are assumed to follow a multivariate Normal distribution with mean vector μ and covariance matrix Σ. The zero components in Σ^{-1} imply the absence of edges in the corresponding graphical model. However, estimating Σ^{-1} by traditional maximum-likelihood estimation for a covariance matrix is not appropriate in a high-dimensional setting. Thus, various algorithms using L_1 regularization to estimate sparse undirected graphical models have been developed in the past decades. In particular, the graphical lasso (glasso) has become one of the most widely adopted approaches due to its computational efficiency. The graphical lasso estimates the inverse of a covariance matrix by maximizing a regularized log-likelihood including a L_1 norm term (Friedman, Hastie and Tibshirani, 2008). As a result, we expect to gain a more comprehensive understanding of the underlying functionality of genes with expression levels that are significantly associated with CYP3A4 activity.

Chapter 2. STATISTICAL METHODS FOR FEATURE SELECTION OF HIGH DIMENSINAL DATA

2.1 FEATURE SELECTION USING REGULARIZED REGRESSION

Suppose that n is the number of observations and p is the number of predictors in the data. Let $\boldsymbol{X}$ be the $n \times p$ data matrix, and let $\boldsymbol{y}$ be the response vector.

Consider the multivariate linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \tag{3.1}$$

where β is the vector of regression coefficients. Recall that the ordinary least squares estimates are obtained by finding β that minimize the residual sum of squares.

$$\| \boldsymbol{y} - \boldsymbol{X}\beta \|^2 \tag{3.2}$$

However, for a high-dimensional dataset (i.e. $p > n$), the least squares regression coefficients can be highly variable due to there being no unique solution for β that minimizes the residual sum of squares.

On the other hand, Lasso (Tibshirani, 1996) involves solving

$$\min\{\| y - X\beta \|^2 + \lambda \sum_j^p |\beta_j|\} \tag{3.3}$$

where λ is a non-negative tuning parameter that controls model complexity. This is equivalent to minimize the sum of residual squares plus an L_1 penalty on the regression coefficients. When $\lambda = 0$, we get the ordinary least square. As λ increases, more coefficients will shrink to zero,

such that a sparser model is obtained. In this study, an optimal value of λ was determined using 10-fold cross-validation from a fine grid of λ values.

Ridge regression (Hoerl and Kennard, 1970) estimates the regression coefficient by involving a L_2 penalty

$$\min\{\| y - X\beta \|^2 + \lambda \sum_j^p \beta_j^2\} \quad (3.4)$$

where λ is a non-negative tuning parameter that shrinks the coefficient estimates. Similar to lasso, when $\lambda = 0$, then ridge regression is equivalent to least squares. As λ increases, the L_2 penalty term increases in magnitude such that coefficients shrink towards zero, but no coefficients are set to exactly zero.

However, two shortcomings of lasso are that (1) when $p > n$, lasso can only pick at most n variables, and (2) lasso tends to pick only one variable from among a group of correlated variables. To address these problems, the elastic net (Zou and Hastie, 2005) was proposed as a compromise between ridge and lasso regression. Elastic net involves finding β that minimizes

$$\| y - X\beta \|^2 + \lambda_2 \sum_j^p \beta_j^2 + \lambda_1 \sum_j^p |\beta_j| \quad (3.5)$$

Equivalently, the algorithm finds β that minimizes

$$\| y - X\beta \|^2 \quad (3.6)$$

subject to the constrain that

$$\alpha \| \beta \|^2 + (1 - \alpha) \| \beta \|_1 \leq t, \text{ where } \alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1} \quad (3.7)$$

where $0 \leq \alpha \leq 1$ is the penalty weight. For α = 0, elastic net is equivalent to lasso regression, and when α = 1, elastic net performs similar to the ridge regression. The L_1 regularization generates sparse model and the L_2 term encourages group effects and stabilizes the L_1 regularization path. Therefore, the result of the elastic net regularization combines the effects of the ridge and lasso regularization to select correlated predictors and allows to select more predictors than the number of observations.

2.2 Constructing Gene Networks

Suppose that $\boldsymbol{X}$ is an $n \times p$ data matrix that follows a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ. Let $\Theta = \Sigma^{-1}$ and S be the covariance matrix of the data. Then, Θ can be estimated by maximizing the log-likelihood with sparsity-inducing L_1 penalty as follow:

$$\log\det\Theta - \mathrm{tr}(S\Theta) - \lambda \| \Theta \|_1 \tag{3.8}$$

where "tr" is the trace which is the sum of the elements on the matrix diagonal, $\| \Theta \|_1$ is the L_1 norm which is the sum of the absolute value of the element of Σ^{-1}, and λ is a non-negative tuning parameter that controls network sparsity. Similar to lasso regression, when $\lambda = 0$, we get the usual maximum likelihood estimate. As λ increases, more coefficients will be shrunken to zero such that we get a sparser graph. Note that the problem in Equation (3.8) is convex, and for a fixed λ, glasso can solve it efficiently by using a block-coordinate descent algorithm (Friedman, Hastie and Tibshirani, 2008).

The optimal λ can be chosen by cross-validation, Bayesian information criterion (BIC) or other methods [(Friedman, Hastie and Tibshirani, 2008), (Foygel and Drton, 2010)]. In this study, the optimal λ is chosen by the StARS approach (Liu, Roeder and Wasserman, 2010). Let $\Lambda = \frac{1}{\lambda}$

so that a smaller value of Λ corresponds to a sparser graph. When $\Lambda = 0$, then the graph is empty graph with no edges. The StARS method chooses Λ based on stability. As Λ increases, the variability of the graph increases while the stability decreases. The optimal Λ is the point that the graph becomes variable as measured by the stability. The StARS approach tends to "over-select" instead of "under-select", which means that it is acceptable to have some false positive but no false negative. In this study, we aim to find interactions between genes, thus it is tolerable that an edge is present between two genes that actually do not interact with each other. It is more difficult to re-discover an edge that is missed at the beginning than eliminate false positive edges by more advanced biological experiments. In addition, the StARS approach performs better, especially in high-dimensional setting. Both the graphical lasso and StARS are conducted by using the "huge" package in R (Zhao, Liu, Roeder, Lafferty and Wasserman, 2013).

Chapter 3. PREDICTING CYP3A4 ACTIVITY FROM RNA-SEQ LIVER BANK DATA

3.1 DATA DESCRIPTION

3.1.1 *Human Liver Bank*

A total of 360 livers from the University of Washington School of Pharmacy Human Tissue Bank and the St. Jude Human Liver Resource were used. The original collection of these tissues for research purposes was approved by the Human Subjects Institutional Review Boards at the University of Washington and St. Jude Children's Research Hospital. All links between archived tissues and the original donors were destroyed to preserve anonymity and facilitate research. Accordingly, the current study was classified as exempt from IRB review. Additional details about the selection of livers and investigator blinding for sample analysis can be found in our previously published study (Shirasaka et al., 2015).

3.1.2 *RNA Isolation*

Liver RNA was isolated and purified using a NucleoSpin® miRNA kit (Macherey-Nagel, Duren, Germany; Clonetech Labs, Mountain View, CA).), according to manufacturer's protocol. Only RNA with RIN greater or equal to 7.0 was submitted for sequencing using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA). Ribosomal RNA was depleted by means of a poly-A enrichment, and first and second strand cDNA syntheses were performed during library construction. Each library was then uniquely barcoded using the Illumina adapters and amplified using a total of 13 cycles of PCR. Library concentrations were then quantified using the Quant-it dsDNA Assay (Life Technologies, Carlsbad, CA). Libraries were subsequently normalized and

pooled based on Agilent 2100 Bioanalyzer results (Agilent Technologies, Santa Clara, CA). Pooled libraries were size selected using a Pippin Prep (Sage Science, Beverly, MA) and then balanced by mass and 9 pooled in batches of 96 with a final pool concentration of 2-3 nM for sequencing on the HiSeq 2500.

3.1.3 *Read Processing and Analysis Pipeline*

The Northwest Genomics Center sequencing lab processing pipeline was used and included the following elements: (1) base calls generated in real-time on the HiSeq or NextSeq instrument; (2) Illumina RTA-generated BCL files converted to FASTQ files; (3) custom scripts developed in-house and used to process the FASTQ files and to output de-multiplexed FASTQ files by lane and index sequence; (4) sequence read and base quality checked using the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/); (5) sequences aligned to hg19 with reference transcriptome Ensembl v67 using Tophat (Kim et al., 2013) followed by mate-fixing; and (6) custom scripts for quality assessment generate metrics. All aligned read data were subject to the following steps: (1) lane level bam data files were merged using the Picard MergeSamFiles tool and suspected PCR duplicates were marked, not removed, in the alignment files using the Picard MarkDuplicates tool (http://broadinstitute.github.io/picard/); (2) local realignment performed around indels, and base quality score recalibration was run using GATK tools (McKenna et al., 2010); (3) variant detection performed with the GATK Unified Genotyper version 2.6.5 (DePristo et al., 2011); (4) aligned data were used for isoform assembly and quantitation with Cufflinks (Kim et al., 2013; Trapnell et al., 2013); genomic features were quantitated with featureCounts (Liao et al., 2014); and (5) gene-specific quantitation data were used for further analysis.

3.2 METHODS

Out of the 360 liver tissue samples collected, 231 had data on both gene expression level and CYP3A4 activity. This subset of 231 livers was then randomly divided into a training set (n = 115) and a test set (n = 116) for prediction. Models were fit using the training set first, and then the predictive ability of the models was evaluated using the test set of livers.

Among all genes in the dataset, 13,232 had at least 200 liver samples for which the fragments per kilobase of transcript per million mapped reads (FPKM) values were greater than 0.5. Only those genes were kept for further analysis.

A *priori*, we hypothesized that age, gender, liver bank site and cytochrome P450 reductase activity may potentially be associated with CYP3A4 metabolic activity but are not in our causal pathway of interest, so these variables were adjusted for in our analyses. We conducted separate linear regression analyses for CYP3A4 metabolic activity with each of the 13,232 genes included as a predictor, as well as the previously mentioned covariates. False Discovery Rate (FDR) was applied to identify significant genes while simultaneously accounting for multiple testing (Benjamini and Hochberg, 1995). FDR is defined as the proportion of discoveries that are false positive. An FDR threshold of 5% was used, which means no more than 5% of the null hypotheses that are rejected are expected to be incorrectly rejected (i.e, are false positives). By controlling the FDR, we are protecting against having too many false positive genes while simultaneously allowing genes that are true positives to have a high probability of being selected.

After pre-screening, the machine learning algorithms were performed with and without FDR correction. Selection of covariates was conducted in two ways: 1) allow the algorithm to pick covariates freely, or 2) all covariates are forced to be included in the prediction model.

Correlations between predicted and actual activity were used to assess prediction performance of the methods, where prediction of activity from gene expression values in the test set was obtained using coefficients that were calculated from the training set. We also present correlations between predicted and actually activity values that were calculated using coefficients from the test set, however, it is important to note that a larger correlation in this setting for the different machine learning prediction models could be due to having a larger number of genes included in the model.

3.3 Prediction Results Using Data from a Previous Study of CYP3A4 Activity

Yang et al. (2010) previously identified a set of genes that were marginally associated with CYP3A4 activity. In this paper, CYP3A4 activity was measured by two substrates, midazolam and testosterone, where these activity measures were designated as CYP3A4M and CYP3A4T, respectively. Age, gender and study site were adjusted for in their analysis, and correlations between a selected set of 5811 transcriptionally active transcripts and the activity measurements of CYP3A4 were assessed. Table 1 shows the top correlated genes from their study.

As shown in Table 3.1, thousands of genes are correlated with each of CYP3A4 activity measurements at $P < 0.05$ level. Note that both CYP3A4M and CYP3A4T activities were significantly and positively associated with the expression of the corresponding coding gene, CYP3A4. Furthermore, CYP3A4 gene was one of the top five correlated genes with CYP3A4M and CYP3A4T activity measurements.

Table 3.1. Summary and top correlated genes with CYP3A4 activity in Yang's study

	CYP3A4M	CYP3A4T
Number of correlated genes at $p<0.05$	3236	2836
P-value for corresponding coding gene CYP3A4	$<5.11*10^{-34}$	$6.66 * 10^{-25}$
Correlation for corresponding coding gene CYP3A4	0.55	0.49
Top correlated genes	LOC285626, THADA, CYP3A64, CYP3A4, CYP3A7	LOC285626, THADA, CYP3A4, CYP3A64, CYP3A7

We then matched those genes having p-value less than 0.05 in our data set. There were 1702 and 1491 significant genes for CYP3A4M and CYP3A4T activity, respectively. We conducted lasso and elastic net algorithms on those genes with our dataset. As mentioned in Section 3.2, we treated the covariates in two ways. Table 3.2 and Table 3.3 summarize the results, respectively.

As shown in Table 3.2, the correlation between measured activity and predicted activity improved by adding the genes that were picked by lasso/elastic net. When we used the genes that marginally associated with CYP3A4M to predict CYP3A4 activity in our dataset, the corresponding coding gene CYP3A4 was picked by both lasso and elastic net. However, CYP3A4 gene was not selected when we used the genes that marginally associated with CYP3A4T. In addition, liver bank site was the only covariate considered to be significant by the machine learning algorithm for when we used the genes that marginally associated with CYP3A4T. Table 3.3 summarizes the result when we forced all the covariates in the model. The prediction performance was also improved by adding the genes that were picked by the machine

learning algorithms. When we used the genes that marginally associated with both CYP3A4M and CYP3A4T, the corresponding coding gene CYP3A4 was not picked by either of the algorithms. However, most of the genes selected were the same for both sets of genes.

Table 3.2. Results of CYP3A4 prediction using genes from Yang's study with free covariates

		Genes marginally associated with CYP3A4M	Genes marginally associated with CYP3A4T
	Number of genes	1702	1491
	Correlation with only covariates	0.167	
Lasso	Number of genes having non-zero coefficient	10	17
	Correlation calculated with training set coefficients	0.286	0.418
	Correlation calculated with test set coefficients	0.642	0.737
	List of picked genes	CYP2B6, **CYP3A4**, UGT2B11, S100A8, RTKN, SOX6, TFAP4, KCNK1, COL6A1, SAT2	CYP2B6, SULT1A2, CYP2A6, CYP2A7, UGT1A5, UGT1A3, PANK1, NTHL1, ACAA1, MTSS1, SCML1, ISG20, VIL1, ACVR2B, CEP57, IL11RA, GBP3
	Any covariates picked	N.A.	Site
Elastic net	Number of genes having non-zero coefficient	17	16
	Correlation calculated with training set coefficients	0.435	0.399
	Correlation calculated with test set coefficients	0.723	0.736
	List of picked genes	CYP2C19, CYP2B6, **CYP3A4**, UGT2B11, FAH, AOX1, PROZ, S100A8, SCAMP5, RTKN, NUCKS1, SLC23A2	CYP2B6, SULT1A2, CYP2A6, CYP2A7, UGT1A5, UGT1A3, PANK1, NTHL1, MTSS1, SCML1, ISG20, VIL1, ACVR2B, CEP57, IL11RA, GBP3
	Any covariates picked	N.A.	Site

Table 3.3. Results of CYP3A4 prediction using genes from Yang's study with forced covariates

		Genes marginally associated with CYP3A4M	Genes marginally associated with CYP3A4T
	Number of genes	1702	1491
	Correlation with only covariates	0.167	
Lasso	Number of genes having non-zero coefficient	6	5
	Correlation calculated with training set coefficients	0.331	0.411
	Correlation calculated with test set coefficients	0.761	0.558
	List of picked genes	CYP2B6, CYP2A6, SULT1A2, UGT1A5, UGT1A3, NTHL1	CYP2B6, SULT1A2, CYP2A6, CYP2A7, UGT1A5, NTHL1
Elastic net	Number of genes having non-zero coefficient	14	6
	Correlation calculated with training set coefficients	0.321	0.398
	Correlation calculated with test set coefficients	0.761	0.558
	List of picked genes	CYP2B6, CYP2A6, CYP2A7, SULT1A2, UGT1A5, UGT1A3, NTHL1, MTSS1, SLC20A2, VIL1, NFIL3, GBP1, ENO3, LRP6	CYP2B6, SULT1A2, CYP2A6, CYP2A7, UGT1A5, NTHL1

3.4 Prediction Result of CYP3A4 Activity Using Whole-genome RNA-seq Data from Liver Samples

In Yang's study, reductase activity was not included in their analysis. We choose to a *priori* include reductase activity in our study. In order to better compare the results of our model to the what was reported in Yang's paper, we performed separate analyses where reductase was

included or excluded as a covariate in the model. The number of genes having significant marginal effects at the 0.05 significance level with CYP3A4 activity were 710 and 901 when we include or exclude reductase as a covariate in the model, respectively. With an FDR = 5% correction, the number of genes that have a significant marginal effect with CYP3A4 activity are 27 and 30, respectively (see Appendix for full lists). Since our lasso and elastic net analyses appropriately handle over-fitting issues, because 10-fold cross-validation was used, we conducted analyses with the machine learning algorithms both with and without FDR correction. In addition, as previously discussed, we treated the covariates in two different ways. Tables 3.4, 3.5, 3.6, 3.7 summarize the results for these different scenarios.

From Tables 3.4 and 3.5 we observed that the results were similar no matter how we treated our covariates, either freely to be selected by the algorithm or forced in the model. The corresponding coding gene, CYP3A4, was the most significant gene, which was as expected. Besides the CYP3A4 gene, other genes, such as CYP2A7, UGT1A2P, CENPN and RP1.21L23.2, were also considered to be important for predicting CYP3A4 activity. In addition, using FDR correction helped improve the prediction performance. Compared to the results in the Yang et al. study, we found that most of the genes selected in two studies were not same, except for CYP3A4, the corresponding coding gene.

From the results reported in Tables 3.6 and 3.7, we observed that reductase activity was highly correlated with CYP3A4 activity. Adding additional genes that were selected by the machine algorithms did not improve the prediction performance much. However, the prediction performance itself was pretty good. The corresponding coding gene, CYP3A4, was still the most significant gene for predicting CYP3A4 activity. Besides CYP3A4 gene, the GAMA gene was selected regardless of whether or not reductase was included in the model. In addition, CYP2A7,

UGT1A2P, CENPN and RP11.21L23.2 were selected only when reductase was not included, whereas AP006285.6, CYP2B6, FAM83A.AS1, CFD and GBP5 were selected only when reductase was included.

Table 3.4. Result of CYP3A4 activity prediction from whole genome RNA-seq data: reductase excluded and free covariates

		Without FDR	With FDR
	Number of genes	901	30
	Correlation with only covariates	0.167	
Lasso	Number of genes having non-zero coefficient	66	3
	Correlation calculated with training set coefficients	0.282	0.418
	Correlation calculated with test set coefficients	0.850	0.737
	List of picked genes	**CYP3A4**, RP11.21L23.2, CTD.3185P2.1, CTAGE9, ATF5, AP001007.1, GBP3…	**CYP3A4**, UGT1A2P, RP11.21L23.2
	Any covariates picked	N.A.	Site
Elastic net	Number of genes having non-zero coefficient	44	6
	Correlation calculated with training set coefficients	0.285	0.399
	Correlation calculated with test set coefficients	0.775	0.736
	List of picked genes	**CYP3A4**, CYP2B6, CYP2A7, UGT1A2P, RP11.21L23.2, CTD.3185P2.1…	**CYP3A4**, CYP2A7, UGT1A2P, CENPN, GZMA, RP11.21L23.2
	Any covariates picked	N.A.	Site

Table 3.5. Result of CYP3A4 activity prediction from whole genome RNA-seq data: reductase excluded and forced covariates)

		Without FDR	With FDR
	Number of genes	901	30
	Correlation with only covariates	0.167	
Lasso	Number of genes having non-zero coefficient	47	5
	Correlation calculated with training set coefficients	0.257	0.418
	Correlation calculated with test set coefficients	0.803	0.536
	List of picked genes	**CYP3A4**, RP11.21L23.2, CTD.3185P2.1, ATF5, AP001007.1, GBP3, NDUFAF6…	**CYP3A4**, UGT1A2P, CENPN, GZMA, RP11.21L23.2
Elastic net	Number of genes having non-zero coefficient	66	7
	Correlation calculated with training set coefficients	0.321	0.419
	Correlation calculated with test set coefficients	0.761	0.597
	List of picked genes	**CYP3A4**, CYP2A7, RP11.21L23.2, CTD.3185P2.1, PPP1R14A, CTAGE9…	**CYP3A4**, SLC39A10, CYP2A7, UGT1A2P, CENPN, GZMA, RP11.21L23.2

Table 3.6. Result of CYP3A4 activity prediction from whole genome RNA-seq data: reductase included and free covariates

		Without FDR	With FDR
	Number of genes	710	27
	Correlation with only covariates	0.5301	
Lasso	Number of genes having non-zero coefficient	76	6
	Correlation calculated with training set coefficients	0.057	0.507
	Correlation calculated with test set coefficients	0.869	0.720
	List of picked genes	**CYP3A4**, GBP3, EAF1.AS1, SLC38A9, ATF5, ZNF582, ATG9B …	**CYP3A4**, AP006285.6, CYP2B6, GZMA, FAM83A.AS1, CFD
	Any covariates picked	Reductase, gender, site	Reductase
Elastic net	Number of genes having non-zero coefficient	54	6
	Correlation calculated with training set coefficients	0.408	0.399
	Correlation calculated with test set coefficients	0.810	0.736
	List of picked genes	**CYP3A4**, CYP2A7, GBP3, RP11.21L23.2, EAF1.AS1, CTAGE9, ATF5…	**CYP3A4**, AP006285.6, CYP2B6, GZMA, FAM83A.AS1, CFD, GBP5
	Any covariates picked	Reductase	Reductase, age, site

Table 3.7. Result of CYP3A4 activity prediction from whole genome RNA-seq data: reductase included and forced covariates

		Without FDR	With FDR
	Number of genes	710	27
	Correlation with only covariates	0.5301	
Lasso	Number of genes having non-zero coefficient	55	5
	Correlation calculated with training set coefficients	0.457	0.484
	Correlation calculated with test set coefficients	0.859	0.608
	List of picked genes	**CYP3A4**, CYP2A7, GBP3, EAF1.AS1, ZNF582, ATG9B, BCL2L10…	**CYP3A4**, AP006285.6, GZMA, FAM83A.AS, CFD
Elastic net	Number of genes having non-zero coefficient	69	7
	Correlation calculated with training set coefficients	0.481	0.489
	Correlation calculated with test set coefficients	0.893	0.733
	List of picked genes	**CYP3A4**, CYP2A7, GBP3, EAF1.AS1, PANK2, ZNF582, ATG9B…	**CYP3A4**, AP006285.6, CYP2B6, GZMA, FAM83A.AS1, CFD, GBP5

Figure 3.1 presents the PCA plots for the 901 genes that had marginal effect on CYP3A4 activity. No obvious pattern was observed from the plots, which indicated that there was no obvious clustering of subsets of genes based on expression values.

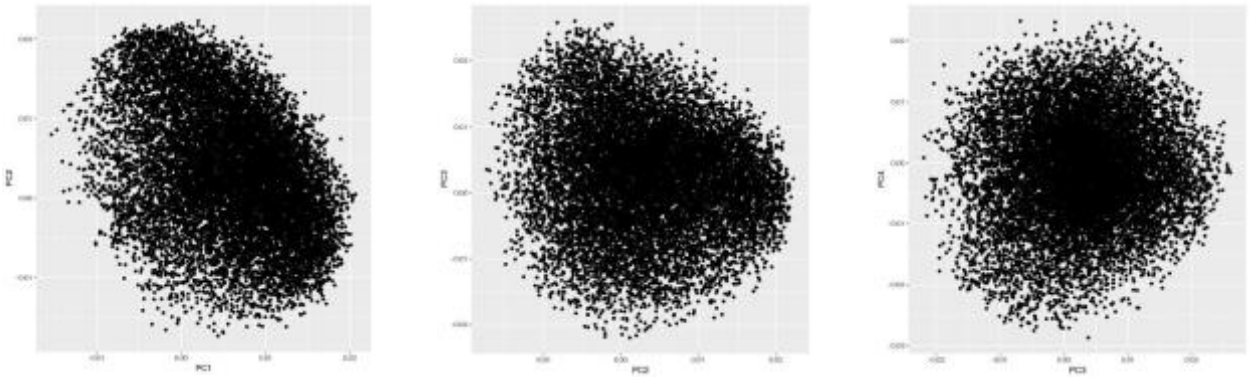


Figure 3.1: PCA plots for genes had significant marginal effect on CYP3A4 activity (from left to right: PC1 v.s. PC2, PC2 v.s. PC3 and PC3 v.s. PC4)

Figures 3.2 and 3.3 show the gene network plots for the significant marginal genes without and with reductase included as a covariate, respectively, after FDR correction. Red points indicated the genes that were picked by the machine learning algorithms. There was a small cluster in both Figures 3.2 and 3.3, indicating that those genes were associated with each other. Most of the genes picked by machine learning algorithms were located inside the cluster.

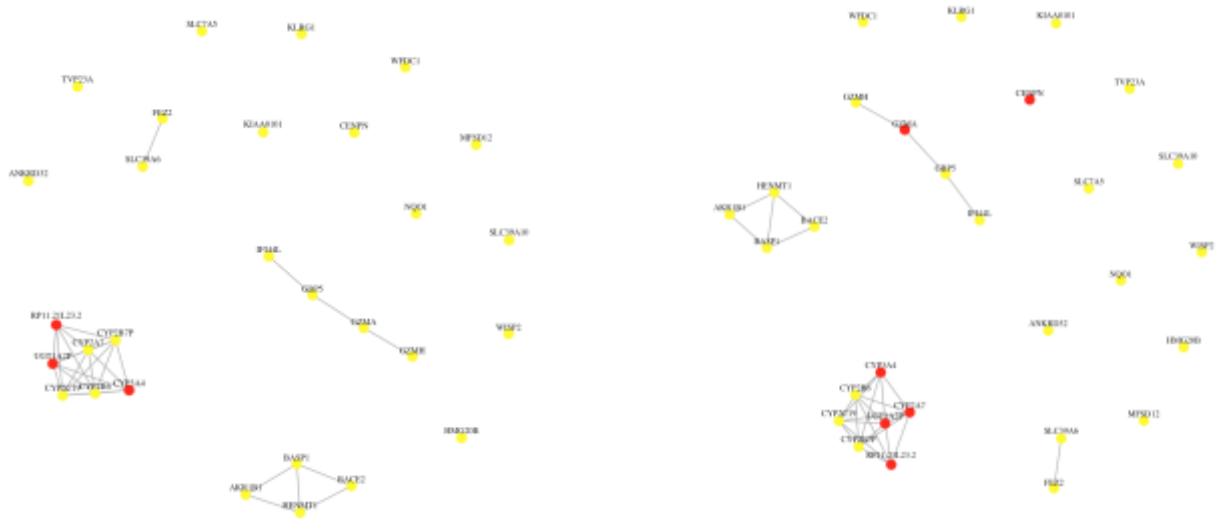


Figure 3.2: Gene network generated by graphical lasso among genes that had significant marginal effect on CYP3A4 activity after FDR=5% correction without reductase (left: lasso, right: elastic net)

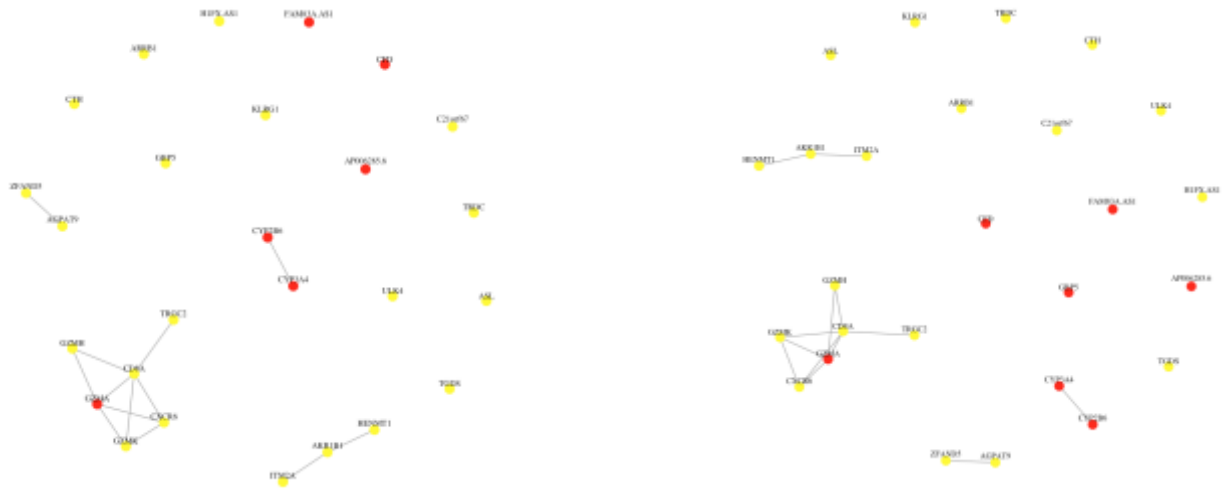


Figure 3.3: Gene network generated by graphical lasso among genes that had significant marginal effect on CYP3A4 activity after FDR=5% correction with reductase (left: lasso, right: elastic net)

Chapter 4. SIMULATION STUDY EVALUATING PREDICTION ACCURACY OF FEATURE SELECTION METHODS

The results given in Chapter 3 indicate that the performance of the machine learning algorithms lasso and elastic net for the prediction of CYP3A4 activity was quite good. However, since this is real data, we don't know the true underlying genes that are driving CYP3A4 activity or how many genes selected by the algorithms were truly associated with CYP3A4 activity. We were interested in assessing the performance of the prediction algorithms if all the truly associated genes (having causal effects) were known. Thus, in this chapter, we present a simulation study and results for assessing the performance of the methods. Under different settings, we compare the performance of lasso and elastic net to a "gold standard" for prediction of a simulated activity phenotype when the causal genes are known. We also evaluate the efficiency of gene selection using the methods.

4.1 SIMULATION PROCEDURE

We first simulated an activity phenotype variable, Y_{sim}, based on the real RNA-seq data from the liver samples that were used to predict CYPR3A4 activity. We then used the same approach described Chapter 3.2 for the analysis of the gene expression data for prediction of the simulated phenotype. In other words, the same RNA-seq/gene expression data was used for prediction of CYP3A4 activity but using a simulated activity phenotype Y_{sim}. By doing so, we are able to assess the performance of the methods when the truth is known.

Y_{sim} was generated based on a linear combination of the expression levels for a set of genes, among which the CYP3A4 gene was always included. The number of genes selected in total, n_{Gene} was set at 5 or 8. Two gene pools were used for gene selection. As for our primary analysis, we selected a total of 13 genes as our gene pool, which are not pairwise statistically associated (or correlated) at 0.05 level, including CYP3A4, COL6A1, RP11.70C1.1, TM2D3, AP001877.1, HMGN4, MRO, USP41, AMN1, RP11.49C24.1, DGCR5, AGPAT6, ZNF614. For a secondary analysis, we allowed for strong associations among genes in the gene pool, so we directly selected genes from those in graphical lasso (Figure 3.2). If some genes were clustered together in the figure, only one of them would be selected. To let each selected gene have relatively the same strength of association with Y_{sim}, the coefficient used in linear combination for each gene was determined to be inversely proportional to the standard deviation of the corresponding gene expression value. The residual error was simulated from a centered Normal distribution. To vary the association between the selected set of genes and Y_{sim}, we generated the residual error with different values for the standard deviation σ used, where σ varied from 1 to 50. To assess the effect that confounding may have on the machine learning algorithms, we also simulated Y_{sim} as a linear combination of a set of potential confounders (covariates), including subjects' age, gender, liver bank site, and reductase activity, in addition to the set of genes.

Using the above data generating procedure, two datasets were generated, with and without covariate effects, for each analysis. For the dataset without covariates effect, we analyzed the data using two different methods, not adjusting for covariates in the analysis and adjusting for covariates. For the simulated phenotype with covariate effects, we only performed the analysis with adjusting for covariates. In addition, when we adjusted for covariates in the

analysis, two different methods were used: (1) fixing all the covariates in lasso and elastic net, and (2) not fixing any of the covariates.

Under each simulation setting, we repeated the procedure 50 times (each with different genes selected), and computed the mean correlation between the predicted values of Y_{sim} and the actual values of Y_{sim}, the mean test error (log10 transformed), the mean number of genes that were selected in total, noted by n_{Total}, and the mean number of genes that were correctly selected, i.e., the selected genes were among the set of genes used to generate Y_{sim}, noted by $n_{Correct}$. In addition, we computed a "gold standard correlation" by using only the true causal genes for the phenotype in the training set and test set. We then summarized the results by plotting the above statistics versus the simulated residual standard deviation σ, which allows us to how the performance of the methods change as a result of increasing noise in the data.

4.2 Simulation Results

The simulation results for the setting where the simulated phenotype data was generated by the five genes that were not pairwise associated at 0.05 level are summarized in Figure 4.1 – Figure 4.3. Figure 4.1 summarizes the results when the phenotype data was generated without covariate effects and no covariates were not adjusted for in the algorithms; Figure 4.2 summarizes the results when the phenotype data was generated without covariate effects but covariates were adjusted for in the algorithm; and Figure 4.3 summarizes the results when the phenotype data was generated with covariate effects and covariates were adjusted for in the algorithm. The case of n_{Gene} = 8 is summarized in Appendix A.3. The simulation results for the case with five causal genes that have strong pairwise associations are summarized in Figure 4.4-Figure 4.6. Similarly, Figure 4.4, Figure 458 and Figure 4.6 summarize the results corresponding to the above three scenarios, respectively.

From Figures 4.1, 4.2 and 4.3, we find that in the case when the data was generated with genes that have weak pairwise association, lasso performed better than elastic net. When σ was small (less than 10 in our setting), both lasso and elastic net can correctly pick all the true genes, and the prediction performance of lasso was very close to the gold standard. However, when σ got larger, the prediction performance got worse and the number of genes that were correctly picked got reduced. Note that, when σ was extremely large, the prediction performance could get better than the gold standard because the algorithms picked many null genes that explain some of the noise. In general, the results with FDR correction provided a slight improvement in prediction than without using FDR correction. In both Figures 4.2 and 4.3, we observe that the results of the two different ways of treating covariates in lasso and elastic net were similar, with a slight better performance when we didn't fix them in the algorithms.

Similarly, from Figure 4.4, 4.5and 4.6, we observe that lasso still performed better than elastic net, though not as much as compared to the above scenarios. The two different ways of treating covariates yield similar results as well. However, when the simulated phenotype data was generated using associated genes with large effects, FDR correction yielded a worse prediction result than without FDR correction. In addition, with regard to picking the correct casual genes for the simulated phenotype, elastic net performed better than lasso, which is not surprising since elastic net allows for group effects.

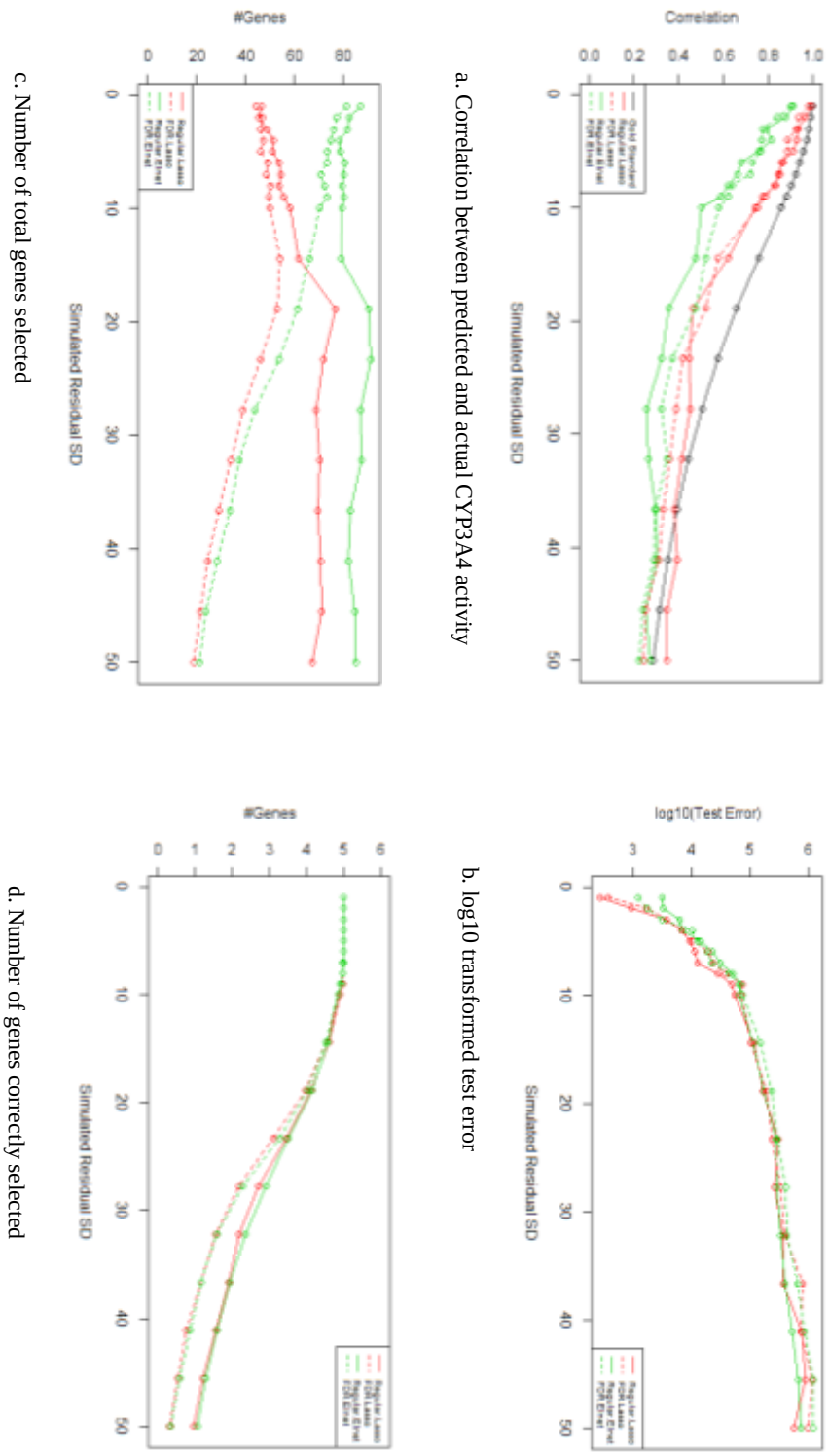
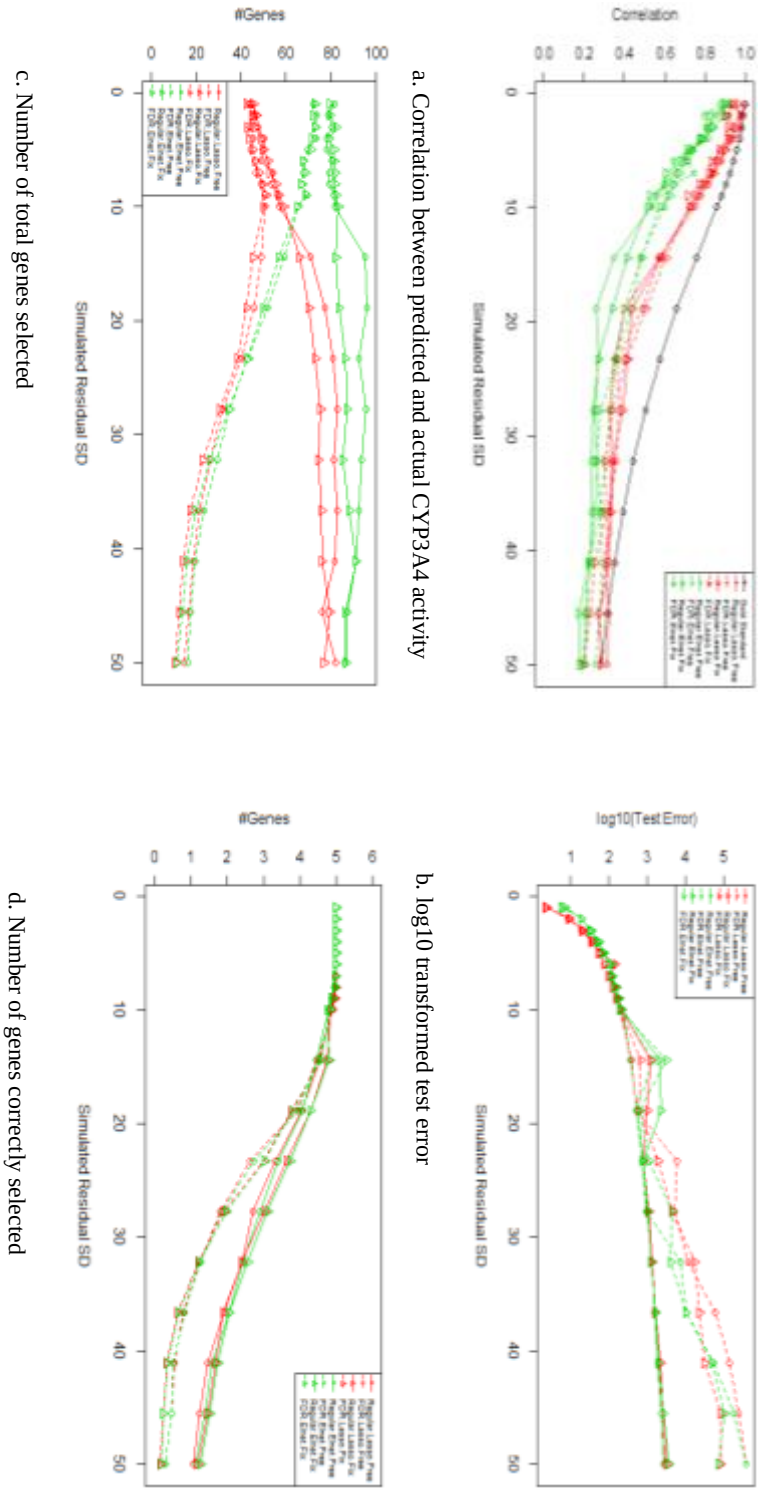


Figure 4.1. Assessing performance of feature selection methods with simulated data where (1) all causal genes have weak pairwise association, (2) there were no covariate effects on the outcome, and (3) no adjustment of covariates in the feature selection methods.



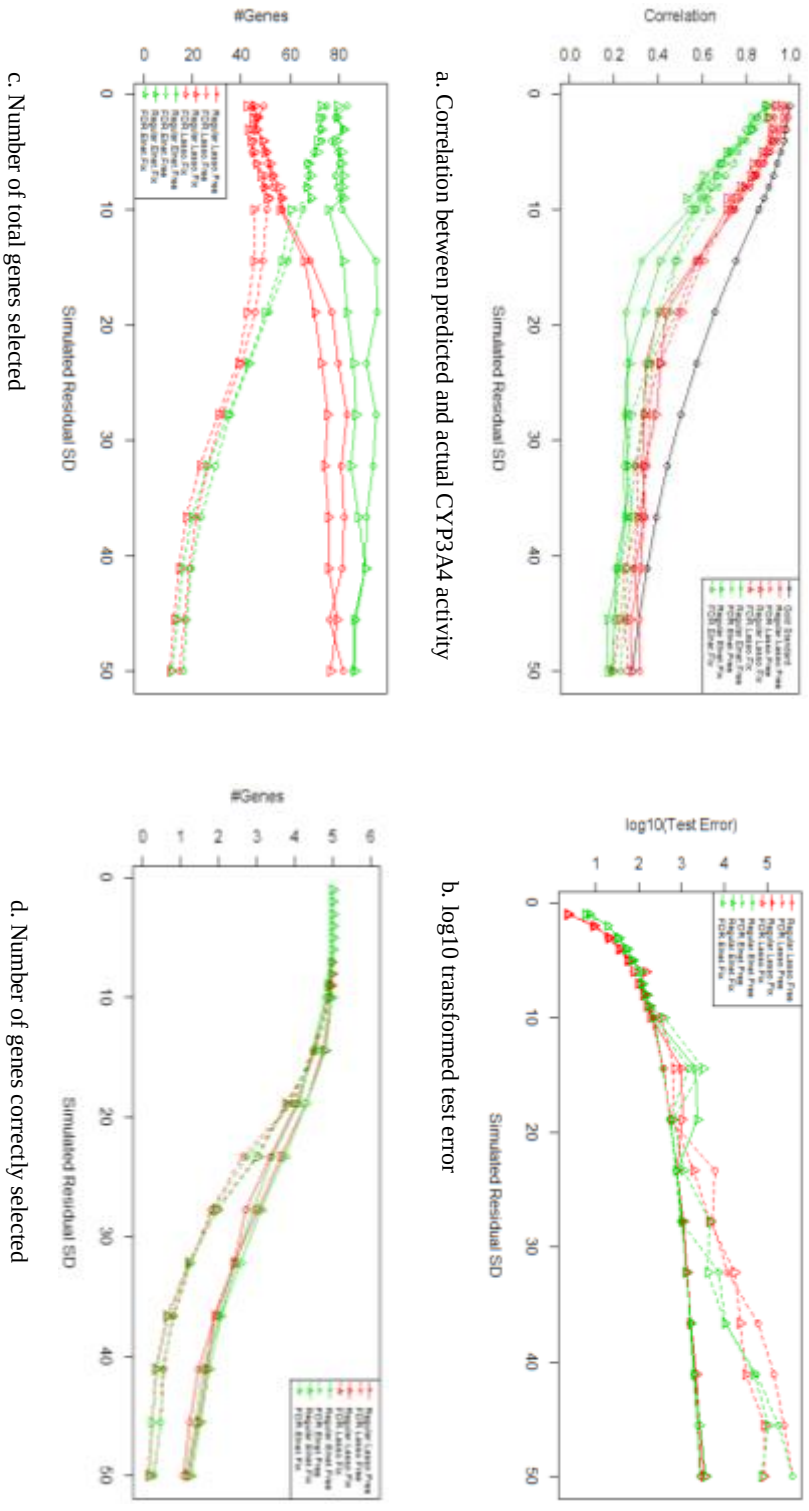
a. Correlation between predicted and actual CYP3A4 activity

b. log10 transformed test error

c. Number of total genes selected

d. Number of genes correctly selected

Figure 4.2. Assessing performance of feature selection methods with simulated data where (1) all causal genes have weak pairwise association, (2) there were no covariate effects on the outcome, and (3) covariates were adjusted for in the feature selection methods.



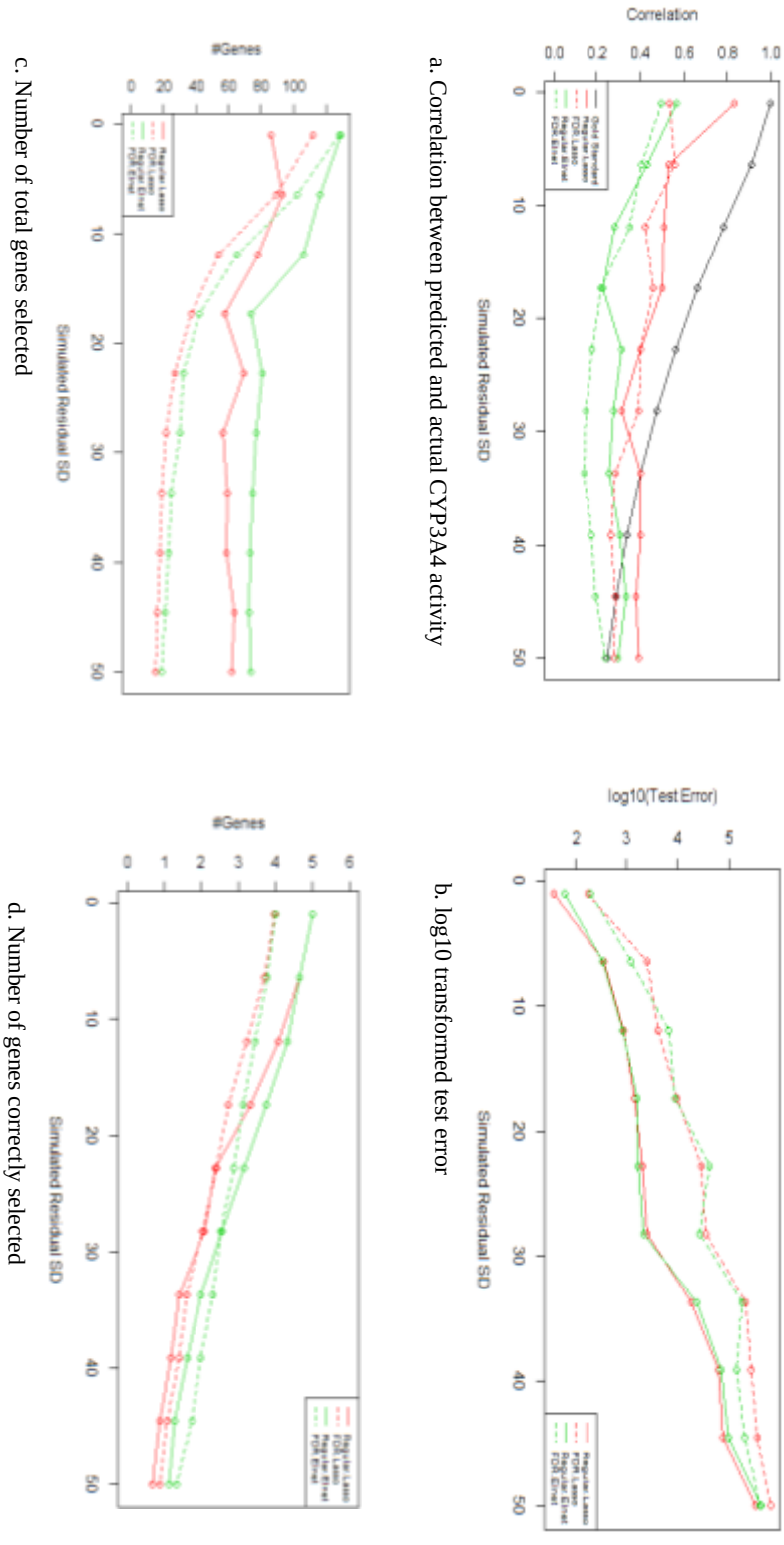
a. Correlation between predicted and actual CYP3A4 activity

b. log10 transformed test error

c. Number of total genes selected

d. Number of genes correctly selected

Figure 4.3. Assessing performance of feature selection methods with simulated data where (1) all causal genes have weak pairwise association, (2) there were covariate effects on the outcome, and (3) covariates were adjusted for in the feature selection methods.



a. Correlation between predicted and actual CYP3A4 activity

b. log10 transformed test error

c. Number of total genes selected

d. Number of genes correctly selected

Figure 4.4. Assessing performance of feature selection methods with simulated data where (1) all causal genes have strong pairwise association, (2) there were no covariate effects on the outcome, and (3) no adjustment of covariates in the feature selection methods.

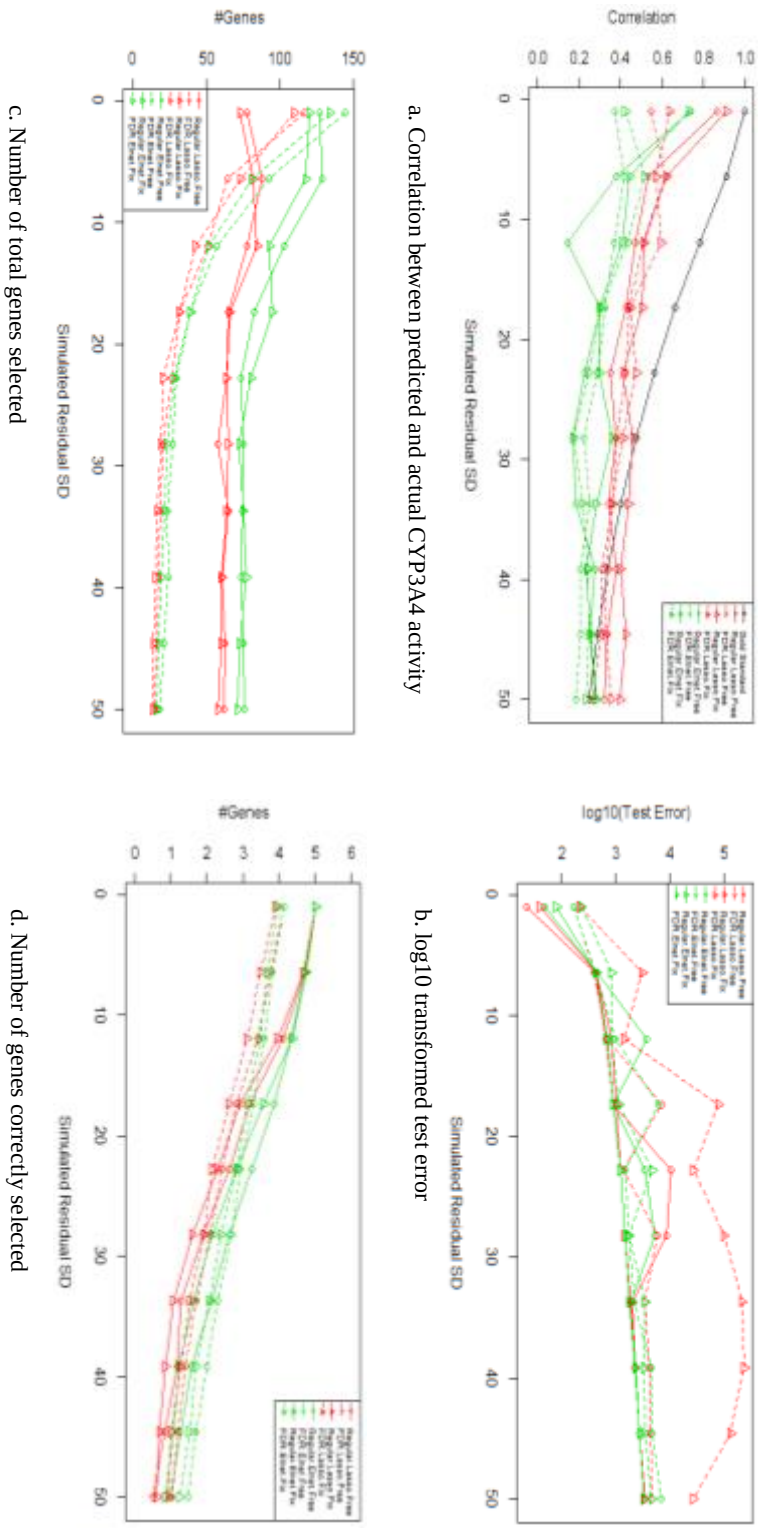


Figure 4.5. Assessing performance of feature selection methods with simulated data where (1) all causal genes have strong pairwise association, (2) there were no covariate effects on the outcome, and (3) covariates were adjusted for in the feature selection methods.

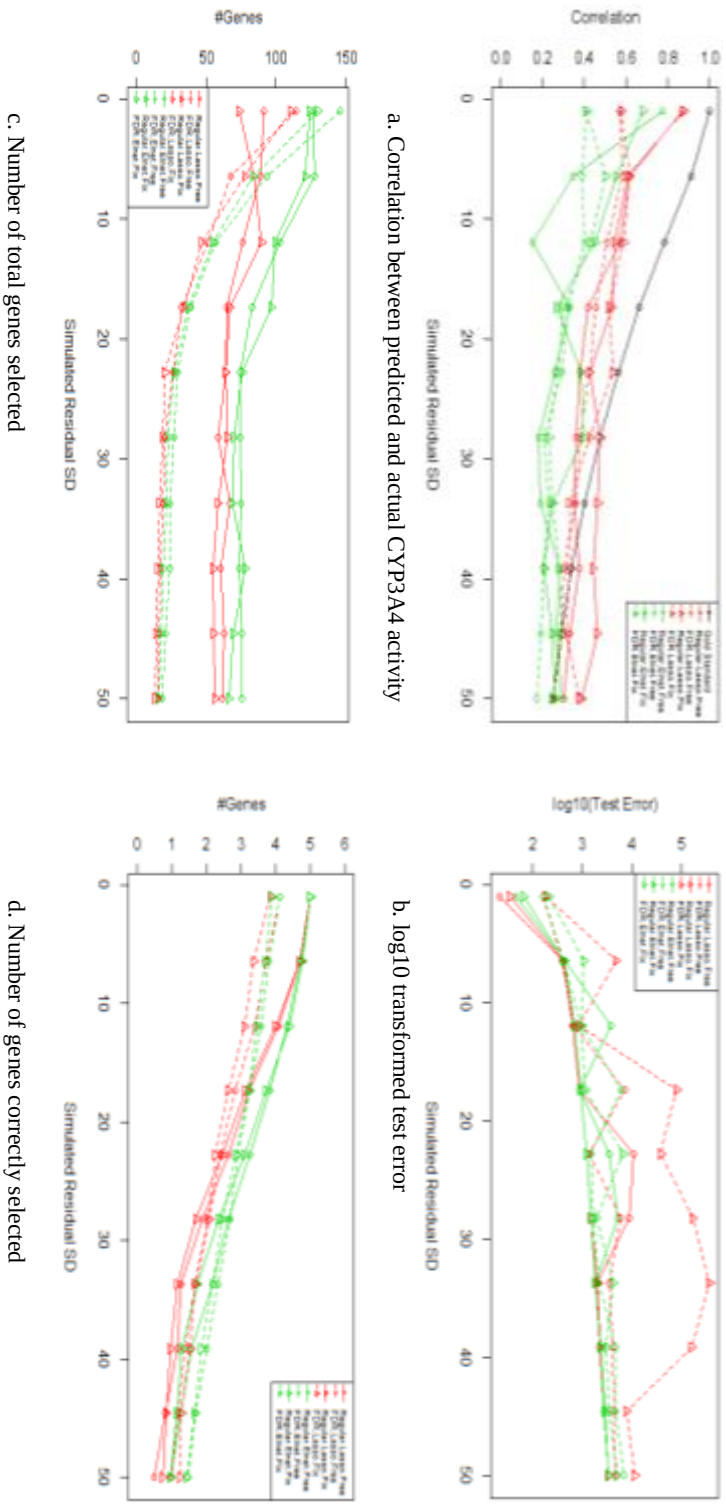


Figure 4.6. Assessing performance of feature selection methods with simulated data where (1) all causal genes have strong pairwise association, (2) there were covariate effects on the outcome, and (3) covariates were adjusted for in the feature selection methods.

Chapter 5. DISCUSSION

In a previous study (Yang et al. 2010), only associations between gene expression and CYP3A4 activity was assessed, but no analysis was performed from the perspective of predicting CYPR3A4 activity. From the results reported by Yang et al., more than three thousand genes had significant associations with CYP3A4 activity, which obviously does not provide much guidance on which genes to focus on in future studies. On the other hand, since genetic factors are known to influence an individual's response to drug metabolism, if a small set of genes that have high prediction efficiency of CYP3A4 activity can be targeted, we may be able to improve prediction of a patient's drug metabolism status, which is essentially for precision medicine.

The prediction performance for both of the machine learning algorithms for CYP3A4 activity was fairly good (the correlation between measured and predicted CYP3A4 activity was around 0.4) when reductase was not included as a covariate and the machine learning algorithms were allowed to pick genes and other covariates freely. In addition, prediction performance of improves from 0.167, when no genes are included as predictors, to 0.4 when genes identified by the feature selection machine learning algorithms are used for prediction of CYP3A4 activity. In this scenario, lasso identified three genes, which are CYP3A4, UGT1A2P, RP11.21L23.2, and elastic net picked six genes, which are CYP3A4, CYP2A7, UGT1A2P, CENPN, GZMA, RP11.21L23.2. As can be seen, the genes identified by elastic net contain all three genes identified by lasso, which indicates that CYP2A7, CENPN and GZMA may have association with CYP3A4, UGT1A2P and RP11.21L23.2, since elastic net allows group effect while lasso does not. From the gene network generated by graphical lasso, it was confirmed that

CYP3A4, CYP2A7, UGT1A2P and RP11.21L23.2 were conditionally dependent with each other. The prediction performance was acceptable, though worse, when covariates are forced in the model as compared to being selected by the machine learning algorithm. The genes selected by the machine learning algorithms were typically the same as when we did not force any covariates to be picked. The lower prediction efficiency might be because covariates that were not associated with CYP3A4 activity were forced into the model, and in previous scenario, only site was considered to be significantly associated with CYP3A4 activity.

When allowing reductase activity to be included as a predictor for CYP3A4 activity, the performance of the machine learning methods significantly improved, and the correlation between predicted and observed CYP3A4 activity was around 0.5 for both lasso and elastic net. However, since reductase activity itself was highly correlated with CYP3A4 activity, including genes identified by the machine learning algorithms did not provide much improvement in the prediction efficiency.

Interestingly, among most analyses there were conducted using the different methods, CYP3A4 and GZMA were always selected by the machine learning algorithms. Thus, these two genes are of high interest to us for further analysis because a better understanding about their association with CYP3A4 activity, as well as reductase activity, could be useful in illuminating other possible pathways for CYP3A4 activity regulation.

From the results of simulation study, we found that our real data example was most similar to the situation when σ was around 20, since we obtained the best correlation around 0.5 in our real data. The mean number of correct genes that can be identified when σ was around 20 was close to 4 when the data was generated by genes with weak correlation, and close to 2 when the data was generated by genes with strong correlation. Besides, if we have a smaller residual

error, then the prediction can be as large as 0.9.

It is important to note that although it was expected that machine learning algorithms can handle the over-fitting issue, the results showed that applying FDR before machine learning algorithms could help improve the prediction performance. With FDR=5% correction for multiple testing, only 30 out of 901 genes were still considered to be marginally significant with CYP3A4 activity, which indicated that there were a lot of noise genes. Thus, lasso and elastic net may not perform well when there was too many noise information.

We should also point out that in this study, we randomly divided the dataset into the training set and the test set, and we set the seed = 1 in the R analysis in order to be able to replicate the results. However, if we changed the seed, the results were sometime not consistent. We have identified two plausible explanations for this. First, the signal to noise ratio may be too low in this data for lasso and elastic net to efficiently handle. Second, the size of the liver bank samples used in the analysis is substantially smaller than the number of genes analyzed.

Finally, it is important to keep in mind that the applications of the machine learning methods discussed in this thesis for predicting CYPR3A4 activity were largely for exploratory analyses. A replication of the results in a future study with an independent dataset will be necessary to have greater confidence that the genes identified by the machine learning algorithms likely play a role in the regulation of CYP3A4 activity.

BIBLIOGRAPHY

[1] Benjamini Y., H. Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

[2] Foyggel R., D. M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *23*, 604-612.

[3] Friedman J, H. T. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432-441.

[4] Hoerl A.E., K. R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

[5] Lamba V., P. J. (2010). Genetic predictors of interindividual variability in hepatic CYP3A4 expression. *J Pharmacol Exp Ther., 332*, 1088-1099.

[6] Liu H., R. K. (2010). Stability approach to regularization selection for high dimensional graphical models. *Advances in Neural Information Processing Systems*.

[7] Ozdemir V, K. W. (200). Evaluation of the genetic component of variability in CYP3A4 activity: a repeated drug administration method. *Pharmacogenetics, 10*, 373-388.

[8] Ortiz de Montellano PR. (2005). *Cytochrome P450: Structure, Mechanism, and Biochemistry.* New York: Luwer Acedemic/Plenum Publishers.

[9] Tibshirani, Robert. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society, 58*(1), 267-88.

[10] Yang X, Z. B. (2010). Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver . *Genome Res., 20*(8), 1020-36.

[11] Zanger U, S. M. (2013). Cytochrome p450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics, 138*(1), 103-141.

[12] Zhao T, L. H. (2013). The huge Package for High-dimensional Undirected Graph Estimation in R. *Journal of Machine Learning Research*.

[13] Zou H, H. T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*, 301-320.

APPENDIX

A.1 List of Genes after FDR Correction

A.1.1 Without reductase

DLG5	WFDC1	SLC7A5	MFSD12	BASP1	TVP23A
CYP3A4	NQO1	ANKRD32	SLC39A10	CYP2B6	IFI44L
FEZ2	CYP2A7	HENMT1	SLC39A6	UGT1A2P	CENPN
BACE2	HMG20B	KLRG1	WISP2	CYP2C19	GZMA
RP11.21L23.2	KIAA0101	CYP2B7P	AKR1B1	GZMH	GBP5

A.1.2 With reductase

ARG2	C21orf67	CYP3A4	CTH	ULK4	ZFAND5
ITM2A	H1FX.AS1	GZMH	AGPAT9	CD8A	TGDS
CXCR6	AP006285.6	CYP2B6	GZMA	ARRB1	KLRG1
ASL	FAM83A.AS1	TRGC2	GZMK	CFD	TRDC
GBP5	AKR1B1	HENMT1			

A.2 Detail Results of Feature Selection Algorithms

A.2.1 Without reductase, free covariates

a. Lasso (without FDR)

List of picked genes:

CYP3A4	RP11.21L23.2	CTD.3185P2.1	CTAGE9	ATF5	AP001007.1
GBP3	NDUFAF6	RHOJ	TTC6	AP000347.2	DIP2C
NID2	AP006285.6	HSPA6	RAPGEF6	GALK1	FBLN5
MMP25	C1orf56	SMAD5	MASP1	MAST4	BX005214.1
ZNF582	ARL4D	WASH5P	AL353671.3	PDCD2L	THSD4
SLC38A9	CRCP	ACVR2A	PKNOX1	RP11.61N20.3	RP11.789C1.1
TSPYL1	RPS6KL1	LINC00238	BX936347.1	ETNPPL	PMS2P4
LRP6	TVP23C	LRRC37B	QDPR	RNA5SP333	GBA3
LRCH1	RP11.1136G11.8	PAX8.AS1	ATG9B	GINS3	PIGQ
PLEKHA8P1	GSTA4	TMEM180	ATXN7L1	CNDP1	LY6E
LIPT1	RP11.181G12.2	LINC00476	PTEN	SEC24B	C9orf116

min lambda: 0.002054117

test error: 0.01099112

b. Elastic net (without FDR)

List of picked genes:

CYP3A4	CYP2B6	CYP2A7	UGT1A2P	RP11.21L23.2
CTD.3185P2.1	PPP1R14A	ATF5	CCBE1	AP001007.1
GBP3	NDUFAF6	RHOJ	TTC6	AP000347.2
DIP2C	NID2	CXCL13	HSPA6	GALK1
FBLN5	MMP25	SMAD5	MASP1	MAST4
ZNF582	ARL4D	WASH5P	SLC11A1	PDCD2L
THSD4	CRCP	ACVR2A	RP11.789C1.1	RPS6KL1
BX936347.1	LRP6	MATN2	RP11.1136G11.8	PAX8.AS1
PLEKHA8P1	TMEM180	CNDP1	LINC00476	

min lambda: 0.01037605
test error: 0.01031523

c. Lasso (with FDR=5%)
List of picked genes:
CYP3A4 UGT1A2P RP11.21L23.2 Site
min lambda: 0.006421186
test error: 0.01041008

d. Elastic net (with FDR=5%)
List of picked genes:
CYP3A4 CYP2A7 UGT1A2P
CENPN GZMA RP11.21L23.2 Site
min lambda: 0.01101753
test error: 0.01018611

A.2.2 Without reductase, fixed covariates
a. Lasso (without FDR)
List of picked genes:

CYP3A4	RP11.21L23.2	CTD.3185P2.1	ATF5
AP001007.1	GBP3	NDUFAF6	RHOJ
TTC6	AP000347.2	DIP2C	NID2
HSPA6	GALK1	MMP25	SMAD5
MASP1	BX005214.1	ARL4D	WASH5P
C5orf24	PDCD2L	THSD4	CRCP
ACVR2A	RP11.789C1.1	ZNF516	SULF2
RPS6KL1	LINC00238	BX936347.1	PMS2P4
LRP6	RNA5SP333	LRCH1	RP11.1136G11.8
PAX8.AS1	BZW1	PIGQ	PLEKHA8P1
RP11.182L21.6	TMEM180	CNDP1	RP11.181G12.2
LINC00476	PTEN	C9orf116	

min lambda: 0.003524473
test error: 0.01047229

b. Elastic net (without FDR)
List of picked genes:

CYP3A4	CYP2A7	RP11.21L23.2	CTD.3185P2.1
PPP1R14A	CTAGE9	ATF5	AP001007.1
GBP3	NDUFAF6	RHOJ	TTC6
AP000347.2	DIP2C	NID2	CSRP2BP
RP11.204M4.2	HSPA6	RAPGEF6	GALK1
MMP25	SMAD5	MASP1	BX005214.1
ARL4D	WASH5P	AL353671.3	PDCD2L
THSD4	SLC38A9	CRCP	ACVR2A
RP11.789C1.1	ZNF516	PKNOX1	SULF2
RP11.61N20.3	RPS6KL1	LINC00238	BX936347.1
ETNPPL	PMS2P4	LRP6	LRRC37B
QDPR	RNA5SP333	LRCH1	BCL2L10
MATN2	RP11.1136G11.8	SFXN2	PAX8.AS1
BZW1	PIGQ	PLEKHA8P1	RP11.182L21.6
GSTA4	TMEM180	ATXN7L1	CNDP1
RP11.181G12.2	LINC00476	PTEN	SEC24B
CTD.3203P2.1	C9orf116		

min lambda: 0.00397373
test error: 0.01056778

c. Lasso (with FDR=5%)
List of picked genes:

CYP3A4	UGT1A2P	CENPN	GZMA	RP11.21L23.2

min lambda: 0.004219398
test error: 0.009869777

d. Elastic net (with FDR=5%)
List of picked genes:

CYP3A4	SLC39A10	CYP2A7	UGT1A2P
CENPN	GZMA	RP11.21L23.2	

min lambda: 0.006818162
test error: 0.009747427

A.2.3 With reductase, free covariates
a. Lasso (without FDR)
List of picked genes:

Reductase	gender	site	
CYP3A4	GBP3	EAF1.AS1	SLC38A9
ATF5	ZNF582	ATG9B	IGLV5.45
BCL2L10	MDP1	TMEM62	TTC6
TMEM88	MTHFS	ARL4D	LINC00476
QDPR	RP5.1125A11.1	GGTLC1	SMAD1.AS1
SMIM10	DENND6A	RP11.49I11.1	HSD17B7
UGT1A3	MAST4	PDSS1	RP11.181G12.2
ZNF419	CSRP2BP	KBTBD4	MMP25

SMAD5	RP11.101E14.2	NUCKS1	FCHSD1
PDCD2L	WASH5P	NDUFAF6	PPP4R4
ACKR2	PDZK1P1	MAT2B	BX005214.1
AC003075.4	PMS2P4	ATAD3C	AP000347.2
CTD.3185P2.1	RHOJ	RAB15	NBPF3
BRWD1	CXCL13	IGLV7.43	GPR97
CCBE1	SNX1	ANKRD10	PITPNC1
MIR17HG	RP11.182L21.6	RP4.798P15.3	RP11.789C1.1
ZNF628	SYT9	TRPM4	PPP1R1C
ACVR2A	NR1D1	CRCP	RP11.1136G11.8

min lambda: 0.001061823
test error: 0.1786205

b. Elastic net (without FDR)
List of picked genes:

Reductase	CYP3A4	CYP2A7	GBP3
RP11.21L23.2	EAF1.AS1	CTAGE9	ATF5
ZNF582	ATG9B	IGL	V5.45 BCL2L10
TTC6	ARL4D	LINC00476	RBFA
QDPR	DENND6A	HSD17B7	UGT1A3
MAST4	PDSS1	RP11.181G12.2	ZNF419
CSRP2BP	KBTBD4	MMP25	SMAD5
RP11.101E14.2	FCHSD1	PDCD2L	WASH5P
PDZK1P1	BX005214.1	AC003075.4	PMS2P4
ATAD3C	AP000347.2	CTD.3185P2.1	RHOJ
CXCL13	CCBE1	ANKRD10	PITPNC1
RP11.182L21.6	KLHL22	HIVEP1	SNTB1
RP11.789C1.1	ZNF628	ACVR2A	CRCP
RP11.1136G11.8	HSPA6		

min lambda: 0.006047322
test error: 0.009210523

c. Lasso (with FDR=5%)
List of picked genes:

Reductase	age	site	
CYP3A4	AP006285.6	CYP2B6	GZMA
FAM83A.AS1	CFD		

min lambda: 0.00397373
test error: 0.009375036

d. Elastic net (with FDR=5%)
List of picked genes:

reductase	age	Site	
CYP3A4	AP006285.6	CYP2B6	GZMA
FAM83A.AS1	CFD	GBP5	

min lambda: 0.008162509

test error: 0.009243564

A.2.4 With reductase, fixed covariates
a. Lasso (without FDR)
List of picked genes:

CYP3A4	CYP2A7	GBP3	EAF1.AS1
ZNF582	ATG9B	BCL2L10	TMEM62
ARL4D	LINC00476	RBFA	QDPR
RP5.1125A11.1	RNF185	SMAD1.AS1	RP11.204M4.2
DENND6A	RP11.49I11.1	MAST4	PDSS1
RP11.181G12.2	ZNF419	KBTBD4	MMP25
SMAD5	RAB33B	RP11.101E14.2	ZMYM6NB
FCHSD1	PDCD2L	WASH5P	ACKR2
PDZK1P1	AC003075.4	PMS2P4	ATAD3C
MORC4	AP000347.2	RHOJ	NBPF3
BRWD1	CXCL13	ANKRD10	PITPNC1
MIR17HG	AC068831.15	RP11.182L21.6	RP11.789C1.1
ZNF628	SYT9	ACVR2A	NR1D1
CRCP	RP11.1136G11.8	SPA6	

min lambda: 0.001821885
test error: 0.009056306

b. Elastic net (without FDR)
List of picked genes:

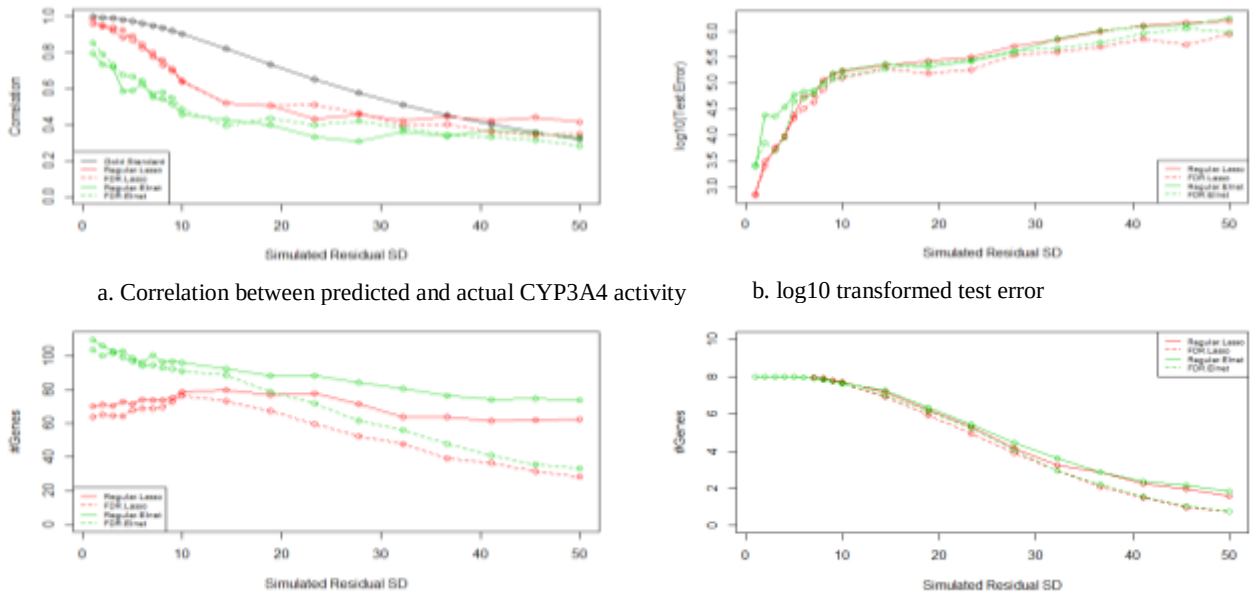
CYP3A4	CYP2A7	GBP3	EAF1.AS1
PANK2	ZNF582	ATG9B	BCL2L10
TMEM62	RNA5SP333	ARL4D	LINC00476
RBFA	QDPR	RP5.1125A11.1	RNF185
GGTLC1	SMAD1.AS1	RP11.204M4.2	DENND6A
RP11.49I11.1	UGT1A3	MAST4	PDSS1
RP11.181G12.2	ZNF419	KBTBD4	MMP25
SMAD5	RAB33B	RP11.101E14.2	ZMYM6NB
NUCKS1	FCHSD1	PDCD2L	WASH5P
ACKR2	PDZK1P1	BX005214.1	AC003075.4
PMS2P4	ATAD3C	MORC4	AP000347.2
CTD.3185P2.1	RHOJ	RHBG	GORASP1
NR3C1	NBPF3	BRWD1	CXCL13
IGLV7.43	ANKRD10	PITPNC1	MIR17HG
AC068831.15	RP11.182L21.6	RFC3	TPSB2
RP11.789C1.1	ZNF628	SYT9	ACVR2A
NR1D1	COQ4	CRCP	RP11.1136G11.8
HSPA6			

min lambda: 0.002944
test error: 0.008841362

c. Lasso (with FDR=5%)
List of picked genes:
CYP3A4 AP006285.6 GZMA FAM83A.AS1 CFD
min lambda: 0.00397373
test error: 0.009006042

d. Elastic net (with FDR=5%)
List of picked genes:
CYP3A4 AP006285.6 CYP2B6 GZMA
FAM83A.AS1 CFD GBP5
min lambda: 0.008667139
test error: 0.008758625

A.3 Additional Results on Simulation Study



a. Correlation between predicted and actual CYP3A4 activity

b. log10 transformed test error

c. Number of total genes selected

d. Number of genes correctly selected

Figure A3.1. Assessing performance of feature selection methods with simulated data where (1) all causal genes (n_{Gene} = 8) have weak pairwise association, (2) there were no covariate effects on the outcome, and (3) no adjustment of covariates in the feature selection methods.