Penalized regression for predicting phenotypes in the Collaborative Cross

by

Jennifer Chen

Senior Honors Thesis

Department of Biostatistics

University of North Carolina at Chapel Hill

May 1, 2020

Approved:

____

____

____

1. Introduction

The Collaborative Cross (CC) are a set of recombinant inbred laboratory mouse strains derived from eight founder strains: 129S1/SvlmJ, A/J, C57BL/6J, NOD/ShiLtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ (Mathes et al., 2011). The CC strains capture a genetic diversity sufficient to produce variation in a large number of phenotypic traits. The combination of genetic and phenotypic diversity allows analyses such as quantitative trait locus (QTL) mapping. Moreover, unlike typical QTL mapping populations, the CC genomes are reproducible, which makes the strains ideal materials for studying responses in isogenic individuals under different interventions (Churchill, 2004).

Since the genetic structures of the CC strains are readily available, the prediction of their phenotypes is particularly useful when the latter are costly to measure. For example, phenotype prediction can help complete a set of CC data that contains several missing phenotype values or identify untested strains that have a desirable range of baseline characteristics as candidate materials for a study.

The genotype at locus m in the genome of an individual i from the CC strains can be represented by the diplotype state, i.e. the pair of founder haplotypes present (Zhang et al., 2014). The diplotype state for individual i at locus m is encoded using the diplotype matrix $\mathbf{D}_i(m)$ such that if the maternally inherited founder haplotype is $j \in 1, \ldots, 8$ and the paternally inherited haplotype is $k \in 1, \ldots, 8$, the entry in the jth row and the kth column of $\mathbf{D}_i(m)$ is $\mathbf{D}_i(m)_{jk} = 1$, while all other entires of $\mathbf{D}_i(m)$ are zeros. The diplotype states cannot be observed directly, but they can be inferred probabilistically through a hidden Markov model (HMM) from the sequencing data. Denote the genotype of n individuals as $\mathbf{G} = \{\mathbf{G}_1, \ldots, \mathbf{G}_n\}$ and genotype of the eight founders as $\mathbf{H} = \{\mathbf{H}_1, \ldots, \mathbf{H}_8\}$, then

$$\mathbf{P}_i(m) = p(\mathbf{D}_i(m)|\mathbf{G}_i, \mathbf{H}) \tag{1}$$

where each entry $\mathbf{P}_i(m)_{jk}$ is the probability that diplotype jk is present for an individual i at locus m (Zhang et al., 2014). When the inheritance at locus m is stable, $\mathbf{P}_i(m) = \mathbf{D}_i(m)$; otherwise, $\mathbf{P}_i(m)$ is affected by genetic marker sparsity, recombination density, and genotyping error (Zhang et al., 2014).

In the CC data, the diplotype states are converted to haplotype dosages. The haplotype dosages for individual i at locus m can be represented as a group of variables $x_i(m)_1, \ldots, x_i(m)_8$, where $x_i(m)_j = \sum_j \mathbf{P}_i(m)_{jk} + \sum_j \mathbf{P}_i(m)_{kj}$.

In genetic studies, linear models are generally used to relate the genotypes to the phenotypes. However, the solution to a simple linear regression with high dimensional genomic data is undefined if the number of predictors exceeds the number of samples. In this case, we are interested in introducing penalization to remove variables with little predictive strength. The lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996) is a penalized regression approach that shrinks the regression coefficients towards zero and performs variable selection by estimating some coefficients to exact zeros. One drawback of lasso is that it disregards any grouping structure in the data. We consider the group lasso (Yuan and Lin, 2006) as an alternative that makes the selection based on the strength of pre-defined groups instead of individual variables. Originally, the group lasso was developed to ensure that when groups of dummy variables are used to encode for categorical factors in the multi-factor ANOVA problems, the variables encoding the same factor are selected or discarded from the model together (Yuan and Lin, 2006). The group lasso is equivalent to lasso when the group sizes are equal to 1.

Given a set of quantitative phenotype data of a subset of genotyped CC strains, we want to train a model that can predict phenotype values of the CC strains for which the data for this phenotype is missing. The objective of this project is to compare the prediction performances and the variable selection consistency of different penalized regression approaches.

2. Methods

2.1 Penalized regression

In a linear model, let the phenotype value of individual i from the CC strains be y_i, then

$$y_i = \mu + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \tag{2}$$

where μ is an intercept, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a p-vector representing p haplotype dosages corresnponding to $m = \frac{1}{8}p$ genetic loci, $\boldsymbol{\beta}$ is a p-vector of effects to be estimated, and $\epsilon_i \sim N(0, \sigma^2)$ is an unobserved random error.

We center the phenotype value so that the observed mean is 0. With n individuals of known genotypes and phenotypes, the ordinary least squares (OLS) estimate of the effect β can be found by solving

$$\hat{\boldsymbol{\beta}}_{OLS} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \tag{3}$$

.

The lasso approach adds a regularization term to the OLS estimate of β such that

$$\hat{\boldsymbol{\beta}}_{Lasso} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_{\ell 1},\ \lambda > 0, \tag{4}$$

where $\|\boldsymbol{\beta}\|_{\ell 1} = \sum_{i=1}^{p} |\beta_i|$ (Tibshirani, 1996). λ is the tuning parameter controlling the scale of the penalties. Large values of λ leads to sparse coefficients and consequently fewer predictors. The lasso estimate is equivalent to the OLS estimate when $\lambda = 0$.

The p predictors belong to m non-overlapping groups such that the predictor index $(1, 2, \ldots, p) = \cup_{j=1}^{j=m} \mathbf{I}_j$. We suppose the cardinality of $\mathbf{I}_j$ is c_j (in our case, $c_j = 8$ for all j). Applying the group lasso,

$$\hat{\boldsymbol{\beta}}_{GroupLasso} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{m} \sqrt{c_j} \|\boldsymbol{\beta}^{(j)}\|_2, \; \lambda > 0, \tag{5}$$

where $\|\boldsymbol{\beta}^{(j)}\|_2 = \sqrt{\sum_{i \in I_j} \beta_{(i)}^2}$ (Yang and Zou, 2014).

2.2 Design of the comparative study

We compared the prediction performances of ridge regression (which introduces shrinkage on the regression coefficients with the penalty term $\lambda \|\boldsymbol{\beta}\|_{\ell 2} = \sum_{i=1}^{p} \sqrt{\beta_i^2}$), lasso, and the group lasso on a dataset available from the control group of a pre-clinical research studying the potential for tolvaptan, a candidate treatment of Autosomal Dominant Polycystic Kidney Disease (ADPKD), to induce liver injuries (Mosedale et al., 2017). The data contain 180 individuals from 45 CC strains; each individual has 61228 predictors at 7641 loci across 20 chromosomes. Three phenotypes were included for analyses: body weights, alanine aminotransferase (ALT) level, and aspartate aminotransferase (AST) levels. The latter two were log-transformed prior to analyses to meet the assumption of normality of the error distribution in linear models. All three phenotype values were averaged for individuals from the same strains, producing 45 samples for testing the models. The lasso and ridge regression were implemented using the *glmnet R* package (Friedman et al., 2010) and the group lasso was implemented with the *gglasso* package (Yang and Zou, 2014).

The tests were carried out via the following leave-one-out cross-validation scheme as recommended (Hastie et al., 2008):

1. A sequence of 20 tuning parameter $\lambda \in [0.01, 10]$ was generated.

2. In each cross-validation cycle, one of the n samples was used as the test sample while the rest were put together as the training samples. A model $f \in \{\text{Ridge, Lasso, Group Lasso}\}$ was trained using all 20 λs from the sequence consecutively.

3. The mean sum of squared error averaged from the n cross-validation cycles was generated for each (f, λ) combination, denoted as $CV(f, \lambda)$,

$$CV(f, \lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i^{(f,\lambda)})^2 \tag{6}$$

where $\hat{y}_i^{(f,\lambda)}$ refers to the estimated phenotype in the i-th cross-validation cycle using model f and tuning parameter λ. For each model f, an optimal λ with the lowest $CV(f, \lambda)$ was identified.

We created a scaled version of $CV(f, \lambda)$ to facilitate the evaluation of a model with any predictors with respect to the intercept-only model, which always predicts $\bar{y}$:

$$CV(f, \lambda)_{scaled} = -log_{10} \frac{CV(f, \lambda)}{CV(f, \lambda)_{intercept-only}} \tag{7}$$

where a $CV(f, \lambda)_{scaled} > 0$ indicates that the inclusion of the predictors in the model helps improve the prediction accuracy. Both scaled and unscaled $CV(f, \lambda)$ were used to compare the model performances.

Another metric that we used to determine the model performances was the coefficient of determination (R^2). R^2 is the proportion of the variance in the responses that is explained by the predictors.

$$\hat{R}^2(f, \lambda) = \max(0, 1 - \frac{CV(f, \lambda)}{Var(\mathbf{y})}) \tag{8}$$

Note that the sum of squared error $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ was approximated by the mean cross-validated sum of squared error $CV(f, \lambda)$. When the predictions of the model exactly match the responses, $\hat{R}^2 = 1$. An intercept-only model will yield $\hat{R}^2 = 0$. Models that have worse performances than the intercept-only model will have negative $\hat{R}^2$ values and were recorded as $\hat{R}^2 = 0$.

3. Results

For each tested phenotype, we report the prediction performances of ridge regression, lasso, and the group lasso using the optimal regularization parameters λs with $CV(f, \lambda)$, scaled $CV(f, \lambda)$, and $\hat{R}^2$. The results are listed in Table 1 to 3.

We often expect applying regularization will help increase the model performances. However, if the regularization is too strong, important predictors may be left out of the model, which makes choosing the λs an essential part of using the penalized regressions. The selection process of the optimal λs are demonstrated in Figure 1 to 6, which show the change in $CV(f, \lambda)$ as λ increases and the number of selected predictor decreases when we used lasso and the group lasso to predict the tested phenotypes in the CC strains. In all figures, the leftmost dotted line indicates the optimal λ we identified for the model, i.e. the λ that yields the lowest $CV(f, \lambda)$. To choose the simplest model whose accuracy is comparable with the best model, we also highlight the largest value of λ such that its $CV(f, \lambda)$ is within one standard error of the minimum for each model as the rightmost dotted line. In some cases, these two λs are equal.

3.1 Phenotype: body weight

Table 1. Comparison of penalized regression methods for body weight prediction in the CC strains (n=45, p=61128)

Phenotype: Body weight					
Model	Best lambda	$CV(f,\lambda)(SD)$	Scaled $CV(f,\lambda)$	$\hat{R}^2$	Number of selected predictors
Ridge	10	9.31E+15 (4.95E+15)	0	0	61077
	0.01	9.12E+15 (4.46E+15)	0.01	0	61077
Lasso	10	13.83 (2.07)	0	0	0
	1.13	13.78 (2.40)	0.002	0	15
Group Lasso	10	13.83 (2.07)	0	0	0

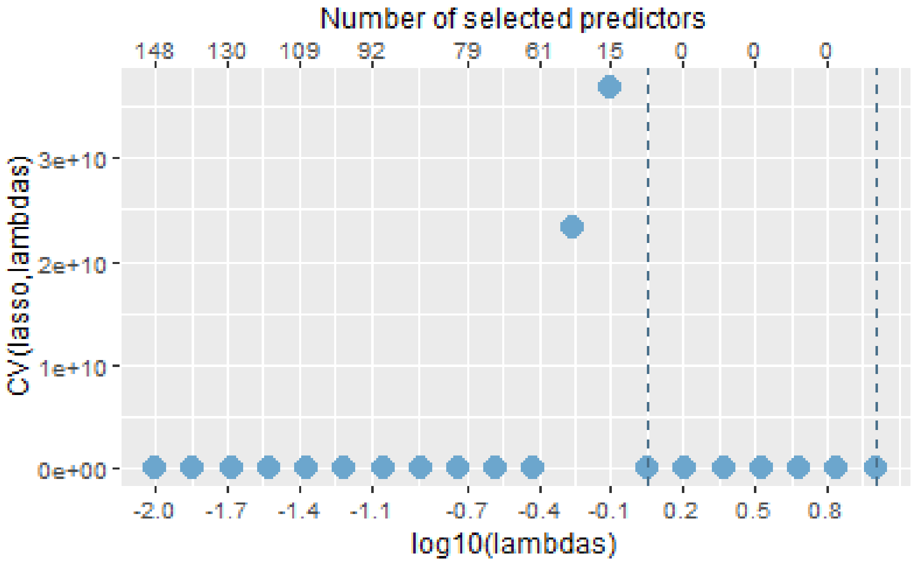

Figure 1: Cross-validated error when using lasso to predict body weight in the CC strains

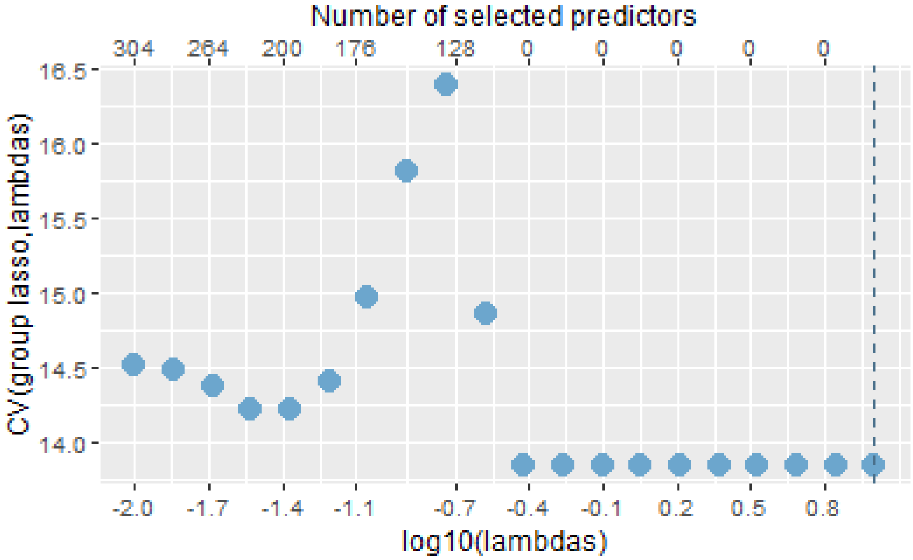

Figure 2: Cross-validated error when using group lasso to predict body weight in the CC strains

3.2 Phenotype: alanine aminotransferase (ALT)

Table 2. Comparison of penalized regression methods for log-transformed ALT level prediction in the CC strains (n=45, p=61128)

Phenotype: log.ALT					
Model	Best lambda	$CV(f,\lambda)(SD)$	Scaled $CV(f,\lambda)$	$\hat{R}^2$	Number of selected predictors
Ridge	2.34	1.69E+13 (6.58E+12)	0.24	0	61077
	0.01	2.22E+13 (8.33E+12)	0.36	0	61077
Lasso	10	0.15 (0.03)	0	0	0
	0.18	0.14 (0.03)	0.04	0.08	2
Group Lasso	10	0.15 (0.03)	0	0	0
	0.02	0.14 (0.03)	0.03	0.04	128

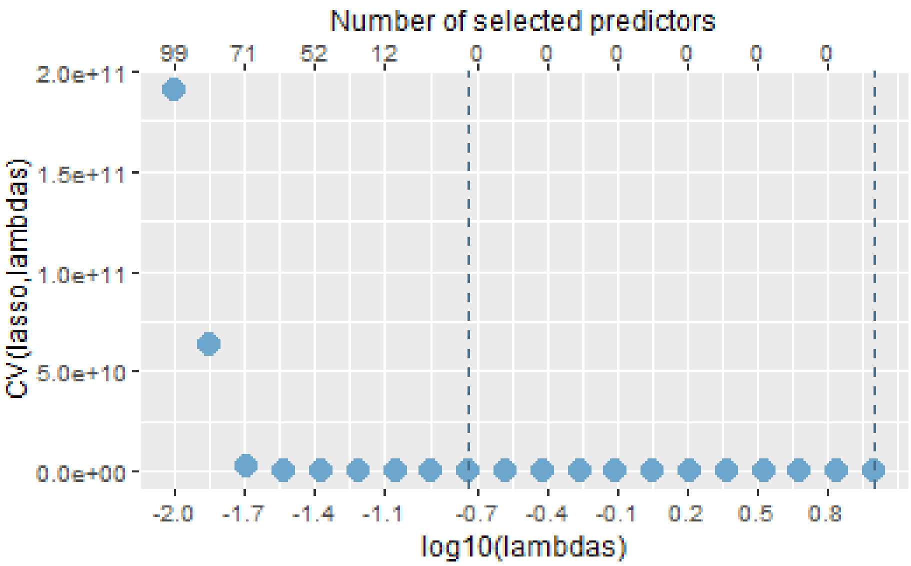

Figure 3: Cross-validated error when using lasso to predict ALT level in the CC strains

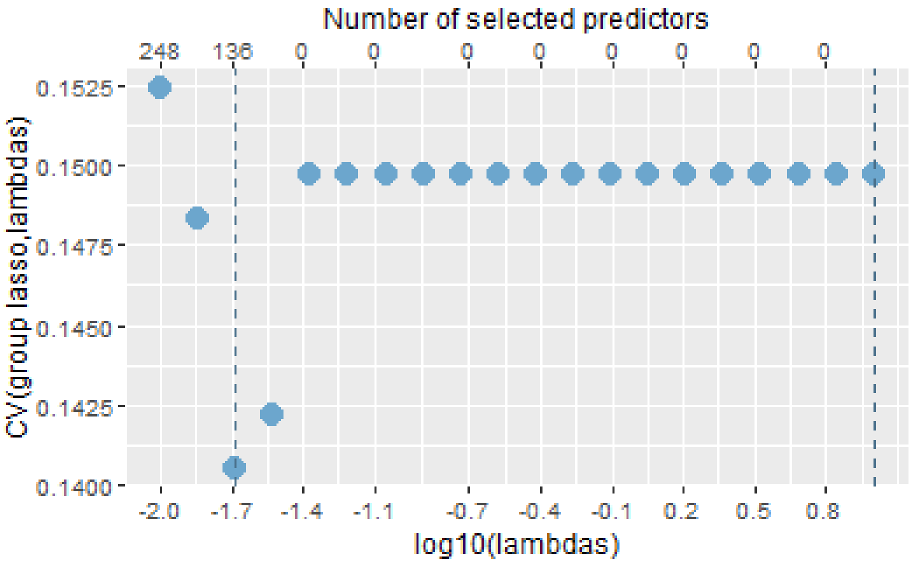

Figure 4: Cross-validated error when using the group lasso to predict ALT level in the CC strains

3.3 Phenotype: aspartate aminotransferase (AST)

Table 3. Comparison of penalized regression methods for log-transformed AST level prediction in the CC strains (n=45, p=61128)

Phenotype: log.AST					
Model	Best lambda	$CV(f,\lambda)(SD)$	Scaled $CV(f,\lambda)$	$\hat{R}^2$	Number of selected predictors
Ridge	10 1.13	2.51E+13 (1.08E+13) 2.29E+13 (1.11E+13)	0 0.04	0 0	61077 61077
Lasso	0.18 0.13	0.13 (0.03) 0.12 (0.02)	0.09 0.05	0.09 0.17	5 17
Group Lasso	0.02 0.01	0.11 (0.03) 0.10 (0.02)	0.13 0.15	0.25 0.28	88 176

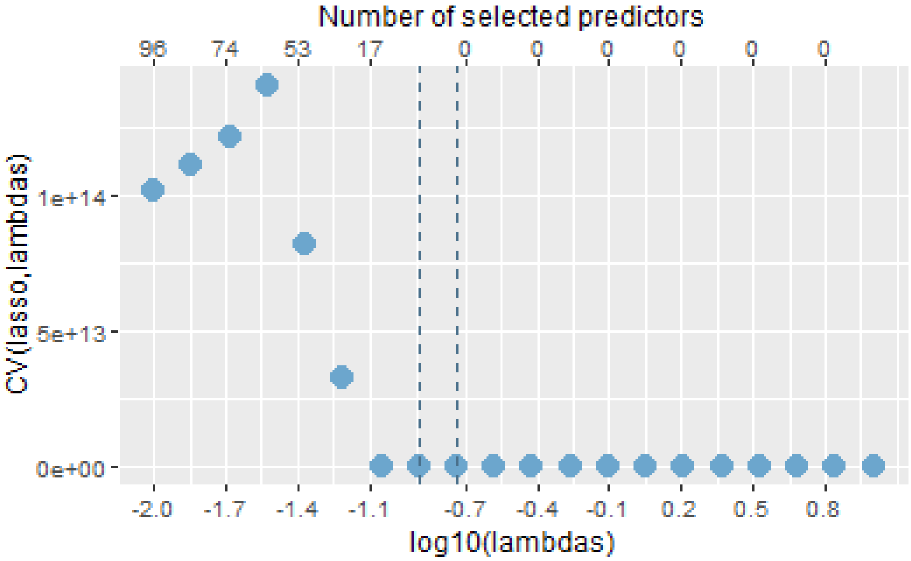

Figure 5: Cross-validated error when using lasso to predict AST level in the CC strains

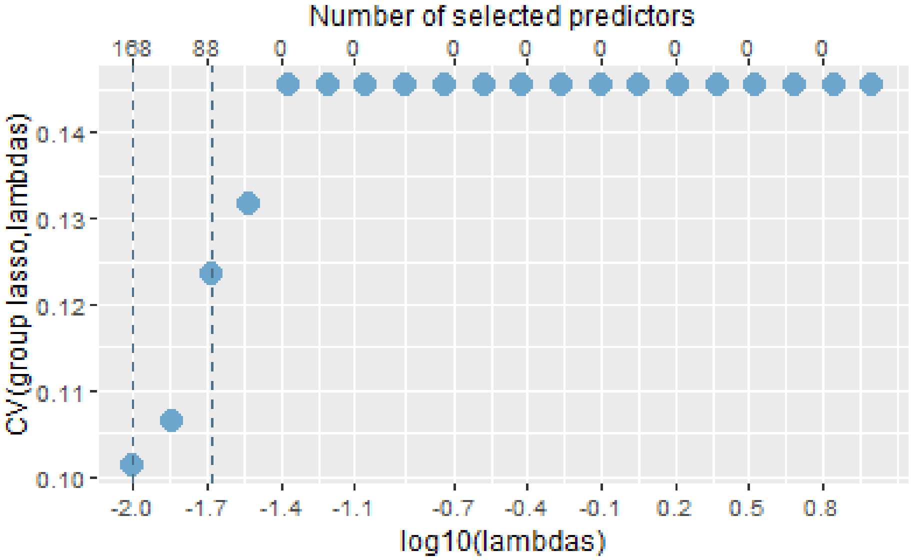

Figure 6: Cross-validated error when using group lasso to predict AST level in the CC strains

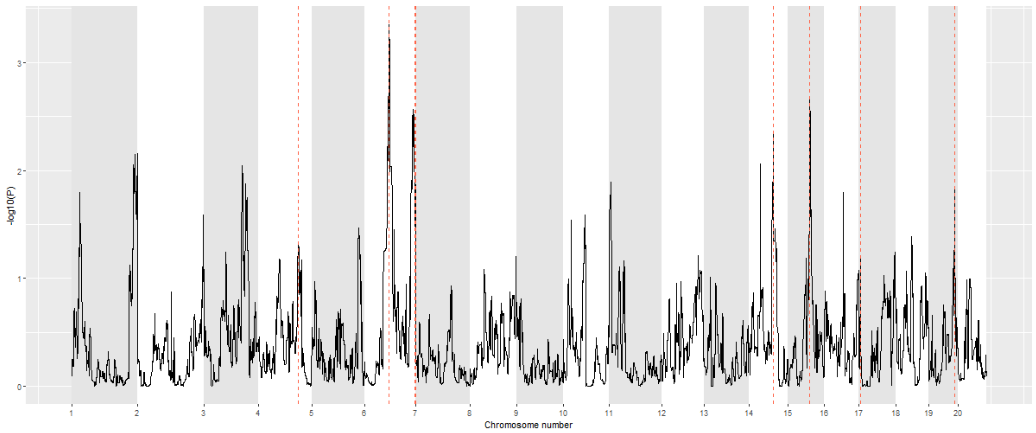

Figure 7: A genome-wide association study for AST level in the CC strains over the predictors selected by the group lasso ($\lambda = 0.02$)

Overall, lasso and the group lasso lead to lower mean cross-validated sum of squared errors when used with the optimal λs. However, most models explain none or little of the variability in the response, except for group lasso achieving $\hat{R}^2 = 0.25$ and 0.28 when used with $\lambda = 0.01$ and 0.02 in the prediction of AST level. We were thus interested in investigating whether the genetic loci selected by this approach would match the loci identified in a genome-wide association study. The strength of association between a genetic locus and a phenotypic outcome was evaluated by the magnitude of the p-value for a linear model that relates the haplotype dosages at the locus to the quantitative phenotype (i.e. the AST level). Figure 7 shows that most genetic loci selected by the group lasso (indicated as dotted lines) align with the loci that are highly associated with the AST level (indicated as the peaks).

4. Discussion

The study compared the prediction performances of three penalized regression methods on the Collaborative Cross data. The relative success of the lasso and the group lasso indicates that the variable selection process significantly increases the prediction performances, which aligns with our knowledge that there exist numerous noise predictors that can be discarded in the CC data. However, despite the success of the group lasso in the prediction of AST level, the two methods failed to find a set of predictors that significantly outperforms the intercept-only model for body weight and ALT level.

Additionally, we investigated the variables selected by different approaches and found that the selected genetic loci do not fully align between the best models of lasso and the group lasso, though there are a lot of overlaps.

To a large extent, the performance of a predictive method depends on the nature of the relationship between the predictors and the response. That most of the compared models did not achieve good performances may be due to the possibility that the linearity assumption is not met for the relationship between the genotype and one or more of the tested phenotypes. In these situations, more complex methods, such as neural networks, can be applied. However, due to our small sample size, such methods tend to overfit the data while linear models are usually more generalizable and suitable for an initial analysis.

In conclusion, we suggest that further investigation is needed to identify the most biologically relevant group of predictors for the tested phenotypes to achieve better predictions.

Bibliography

Churchill, G. (2004). The collaborative cross, a community resource for the genetic analysis of complex traits. *Nat Genet.*, Nov;36(11):1133–7.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*, Jan;33(1):1–22.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Mathes, W. F., Aylor, D. L., Miller, D. R., Churchill, G. A., Chesler, E. J., de Villena, F. P., Threadgill, D. W., and Pomp, D. (2011). Architecture of energy balance traits in emerging lines of the collaborative cross. *Am J Physiol Endocrinol Metab.*, Jun;300(6):E1124–34.

Mosedale, M., Kim, Y., Brock, W. J., Roth, S. E., Wiltshire, T., Eaddy, J. S., Keele, G. R., Corty, R. W., Xie, Y., Valdar, W., and Watkins, P. B. (2017). Candidate risk factors and mechanisms for tolvaptan-induced liver injury are identified using a collaborative cross approach. *Toxicol Sci.*, Apr 1;156(2):438–454.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58, No. 1:267–288.

Yang, G. and Zou, H. (2014). A fast unified algorithm for solving group-lasso penalized learning problems. *Stat Comput*, Stat Comput 25:1129–1141.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68, Part 1:49–67.

Zhang, Z., Wang, W., and Valder, W. (2014). Bayesian modeling of haplotype effects in multiparent populations. *Genetics*, Sept 1;198, No. 1:139–156.