Variable Selection in High-dimensional Regression with Application to Genome-wide Multiple Loci Mapping

Fei Zou

Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina-Chapel Hill
fzou@bios.unc.edu

Genome-wide Multiple Loci Mapping

With the assumption of no interaction and linearity of allelic effects, we employ an additive linear model:

$$y_i = b_0 + \sum_{j=1}^{p} x_{ij} b_j + e_i, \tag{1}$$

where y_i $(i = 1, ..., n)$ is the trait value of individual i, $e_i \sim N(0, \sigma_0^2)$ is the residual error, and x_{ij} is the genotype of the j-th marker of individual i. For example, $x_{ij} = 0, 1, 2$ for genotype AA, AB, and BB.

Note $n < p$, and the majority of the b_js are actually 0. The objective of multiple loci mapping is to identify the correct subset model, i.e., to identify those js, such that $b_j \neq 0$, and estimate those b_j.

Genome-wide Multiple Loci Mapping

Experimental cross of inbred strains

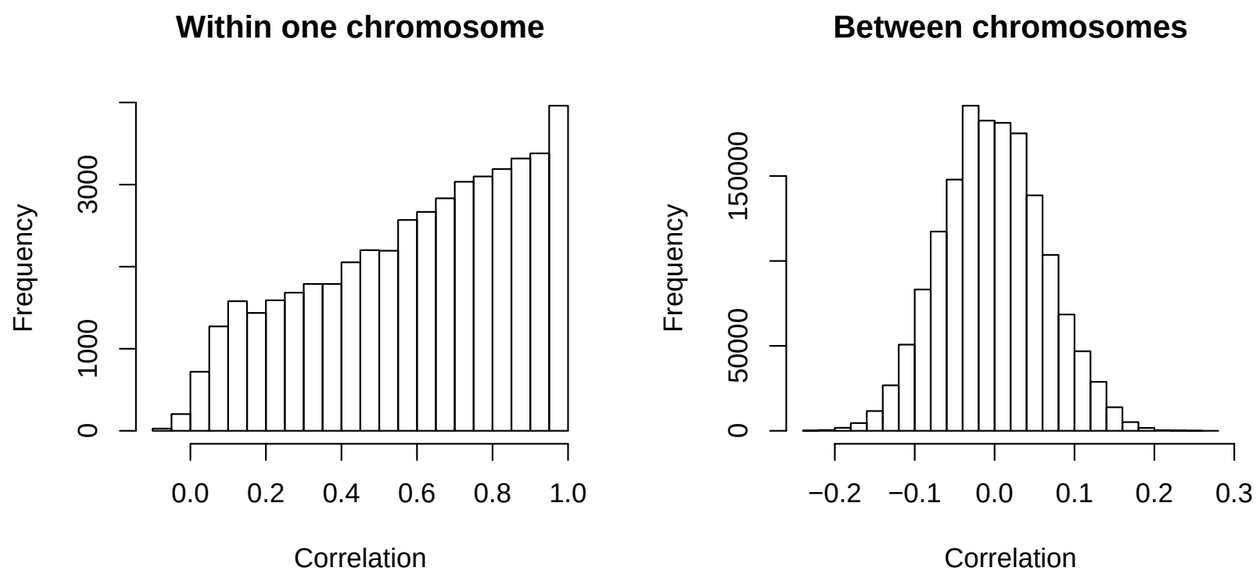
- Thousands of markers, hundreds of samples.
- High correlations among markers' genotype profiles. One marker may be significantly correlated with 2/3 of the markers in the same chromosome.

Genome-wide association studies in human

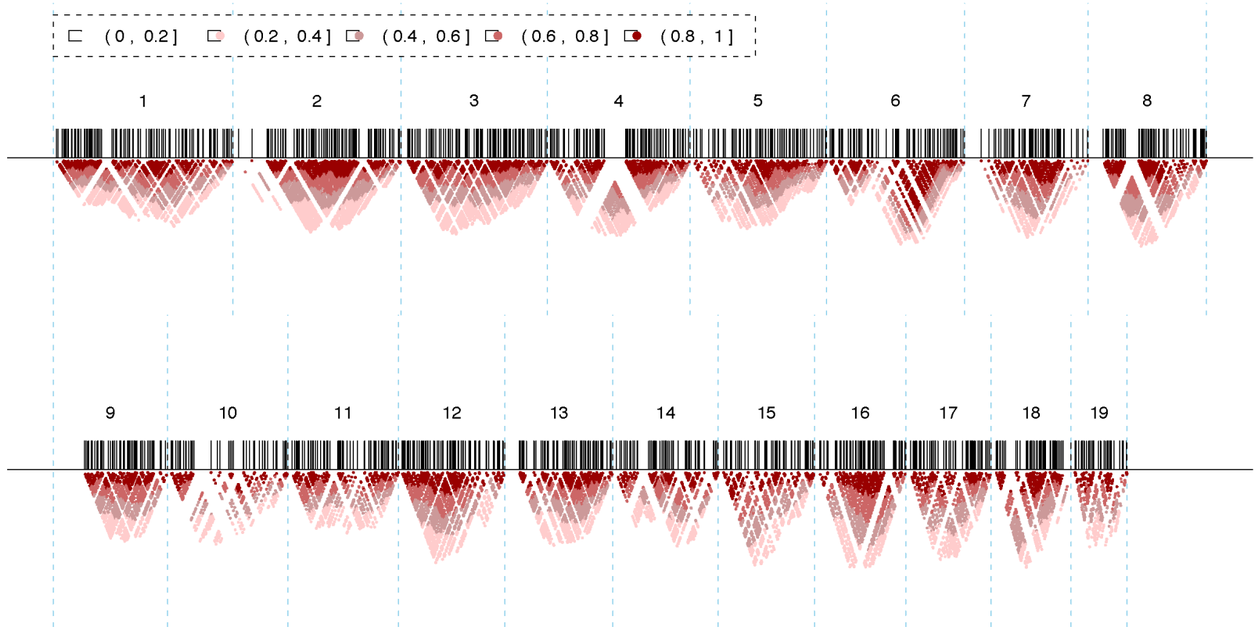
- Millions of markers, thousands of samples.
- Nearby markers have correlated genotype profiles. Length of linkage disequilibrium regions are limited.

We will focus on experimental cross of inbred strains.

Correlations of marker genotype profiles in experimental cross



Correlations of marker genotype profiles in experimental cross



Existing Methods: Marginal regression and step-wise selection

- Marginal regression
  - Multiple testing correction by permutation p-value [Churchill and Doerge 1994]
- Forward regression
  - Multiple testing correction by conditional permutation, conditioning on the markers included in the model [Doerge and Churchill 1996]
- Forward-backward regression
  - Broman et al. proposed a model selection criterion named pLOD (also known as BIC_δ), which is implemented within a framework of forward-backward regression. The threshold of pLOD is estimated by permutations [Manichaikul et al 2008].

Existing Methods: Bayesian shrinkage methods

$$b_j \sim N(0, \sigma_j^2),\ j = 1, ..., p.$$

- Bayesian t [Tipping 2001; Yi and Xu 2008]

$$p(\sigma_j^2|\delta, \tau) = \text{inv-Gamma}(\delta, \tau) = \frac{\tau^\delta}{\Gamma(\delta)}(\sigma_j^2)^{-1-\delta}\exp(-\tau/\sigma_j^2).$$

  The unconditional prior of b_j is a Student's t distribution.

- Bayesian Lasso [Park and Casella 2008; Yi and Xu 2008]

$$p(\sigma_j^2|\kappa^2/2) = \text{Exp}(\kappa^2/2) = \frac{\kappa^2}{2}\exp\left(-\frac{\kappa^2}{2}\sigma_j^2\right).$$

  The unconditional prior of b_j is a double exponential distribution:

$$p(b_j) = \frac{\kappa}{2}e^{-\kappa|b_j|}$$

Existing Methods: Bayesian model selection methods

- The reversible jump MCMC method [Rihard and Green 1997]
- The stochastic search variable selection (SSVS) method [George and McCulloch 1993]

One example is the composite model space approach (CMSA) implemented in R/qtlbim [Yi 2004].

Existing Methods: Penalized regression

Adaptive Lasso extends Lasso by using different penalization parameters for different regression coefficients and has variable selection consistency [Zou 2006]. However, the adaptive Lasso requires consistent initial estimates of the regression coefficients, which are generally not available when $n < p$.

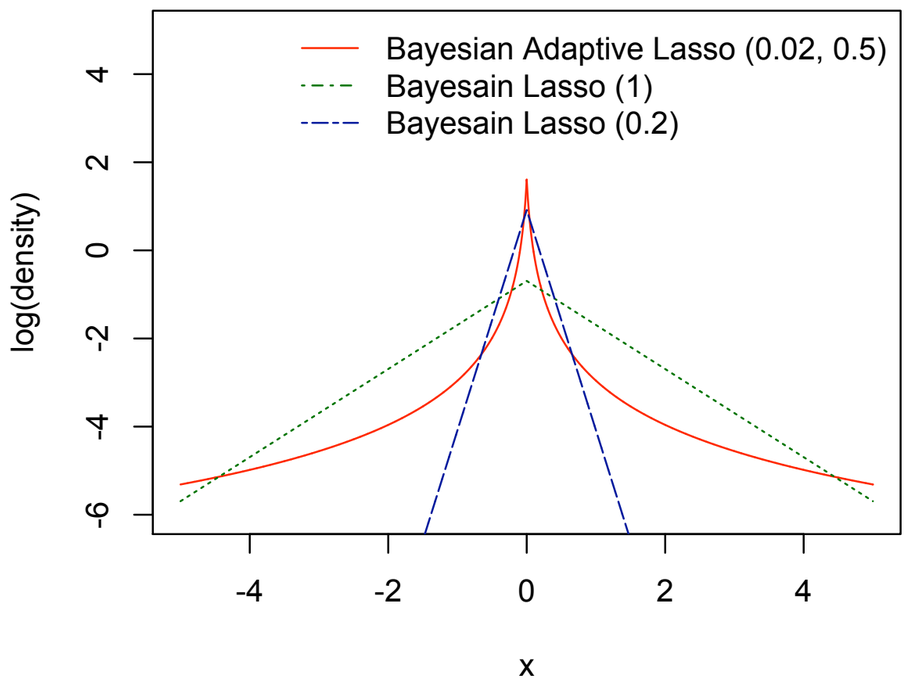
A Bayesian interpretation of the adaptive Lasso

$$p(b_0) \propto 1,\ p(\sigma_0^2) \propto 1/\sigma_0^2,$$
$$p(b_j|\kappa_j) = \frac{1}{2\kappa_j} \exp\left(-\frac{|b_j|}{\kappa_j}\right),$$
$$p(\kappa_j|\delta,\tau) = \text{inv-Gamma}(\kappa_j;\delta,\tau) = \frac{\tau^\delta}{\Gamma(\delta)}\kappa_j^{-1-\delta}\exp\left(-\frac{\tau}{\kappa_j}\right),$$

where $\delta > 0$ and $\tau > 0$ are two hyperparameters. The posterior distribution is sampled by a Gibbs sampler.

A similar approach has been proposed [Griffin and Brown 2007], where $p(\kappa_j^2|\delta,\tau) = \text{inv-Gamma}(\kappa_j;\delta,\tau)$. An efficient implementation is available as HyperLasso [Hoggart et al 2008].

Compare the Bayesian Lasso and the Bayesian Adaptive Lasso



Unconditional prior for the BAL: $f(x;\tau,\delta) = \frac{\tau^{\delta}\delta}{2}(|x|+\tau)^{-1-\delta}$.
Unconditional prior for the Bayesian Lasso $f(x;\kappa) = \frac{1}{2\kappa}\exp(-|x|/\kappa)$.

Iterative Adaptive Lasso (IAL)

Because no point mass at zero is specified in the Bayesian shrinkage methods (including the BAL), the samples of the regression coefficients would not be exactly zero, so that the Bayesian shrinkage methods do not automatically select variables.

However, if we look for the mode of the posterior distribution, it could be exactly zero. This leads to the following ECM algorithm [Meng and Rubin 1993]: the iterative adaptive Lasso.

1. Initialization: We initialize $b_j (0 \leq j \leq p)$ with zero, initialize σ_e^2 by variance of y, and initialize $\kappa_j (1 \leq j \leq p)$ with $\tau/(1+\delta)$.
2. Conditional Maximization (CM) step:
   (a) Update b_0 by its posterior mode,

$$b_0 = (1/n) \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} b_j \right).$$

(b) For $j = 1, ..., p$, update b_j by its posterior mode (see Web Appendix B),

$$\begin{cases} b_j = 0 & \text{if } -\sigma_j^2/\kappa_j \leq \bar{b}_j \leq \sigma_j^2/\kappa_j \\ b_j = \bar{b}_j - \sigma_j^2/\kappa_j & \text{if } \bar{b}_j > \sigma_j^2/\kappa_j \\ b_j = \bar{b}_j + \sigma_j^2/\kappa_j & \text{if } \bar{b}_j < -\sigma_j^2/\kappa_j \end{cases},$$

where

$$\sigma_j^2 = \frac{\sigma_e^2}{\sum_{i=1}^n x_{ij}^2},\ \bar{b}_j = \left(\sum_{i=1}^n x_{ij}^2\right)^{-1} \sum_{i=1}^n x_{ij} \left(y_i - b_0 - \sum_{k \neq j} x_{ik} b_k\right)$$

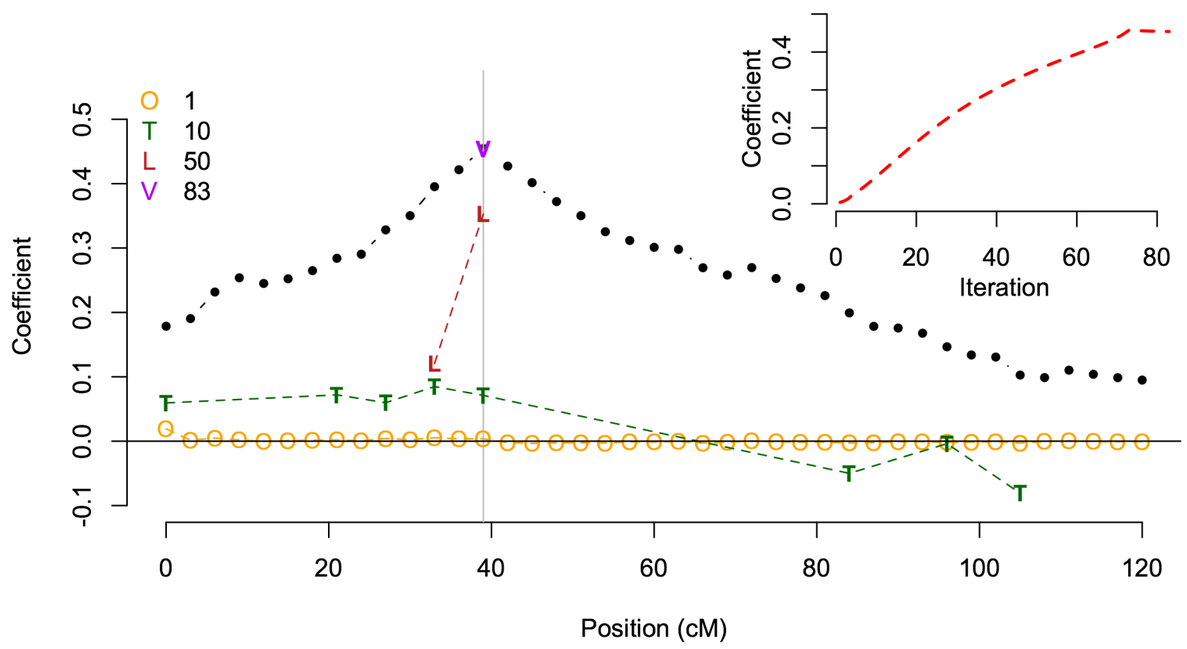
3. Expectation (E) step:

   With the updated b_j's, recalculate the residual sum of squares, `rss`, and

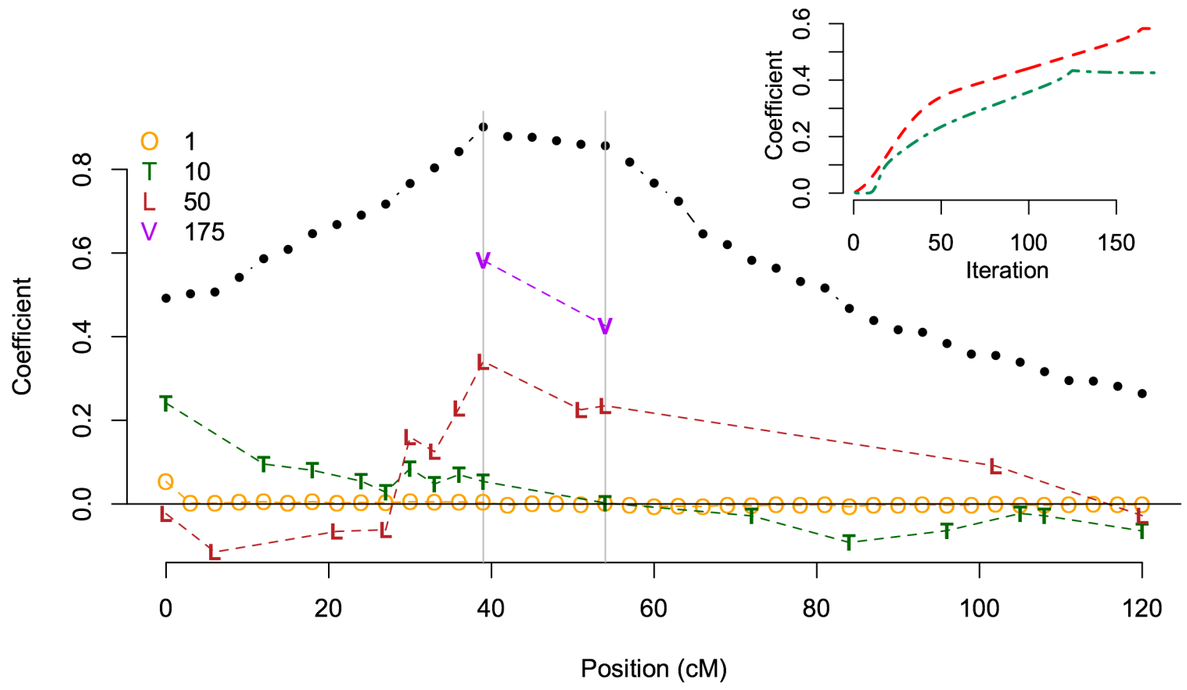
   (a) Update σ_e^2: $\sigma_e^2 = \texttt{rss}/n$.

   (b) Update κ_j: $\kappa_j = (|b_j| + \tau)/(1 + \delta)$.

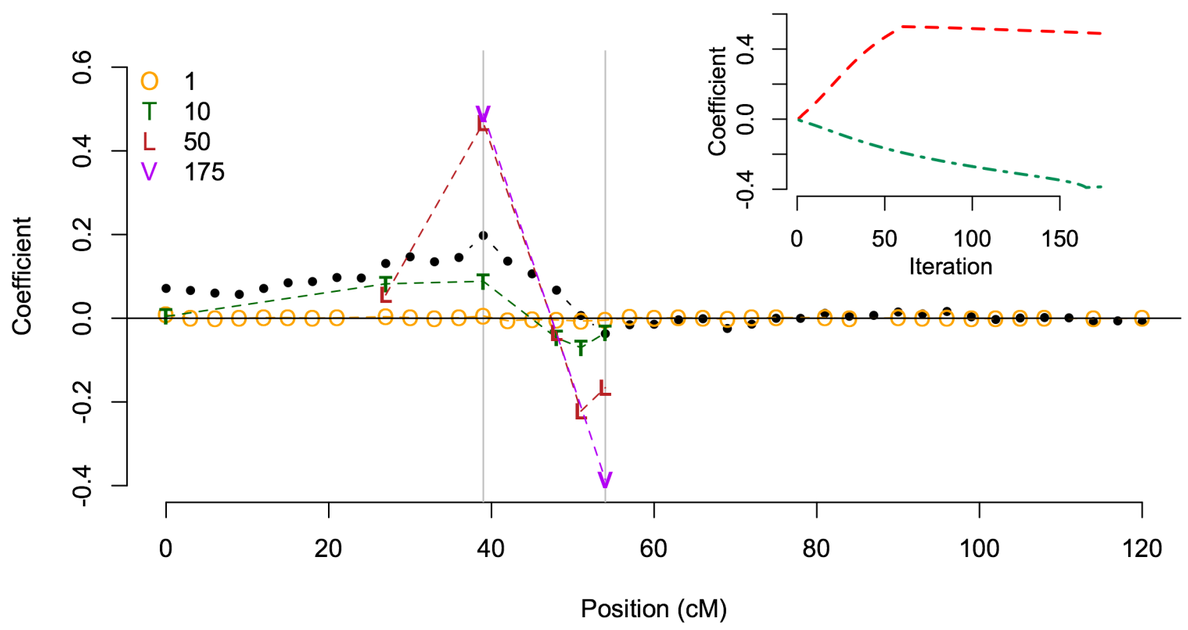
Example of IAL updates: Independent QTL



Example of IAL updates: QTL linked in coupling



Example of IAL updates: QTL linked in repulsion



Iterative Adaptive Lasso (IAL), how to choose δ and τ

Theorem 1. Assume the penalization parameters of the IAL satisfy $(1+\delta)/\tau = O(n^{1/2+d})$, where $0 < d < 1/2$. Denote the coefficient estimates in the t-th iteration as $\hat{\mathbf{b}}^{(t)}$. Let $\mathbf{X}_{-j}$ be $\mathbf{X}$ without the j-th column and let $\tilde{\mathbf{b}}_{-j}^{(t+1)}$ be the coefficient estimates (expect b_j) before estimating $\hat{b}_j^{(t+1)}$.

(i) If $\hat{b}_j^{(t)} = 0$ and $\mathbf{x}_j \perp \mathbf{y} | \mathbf{X}_{-j}\tilde{\mathbf{b}}_{-j}^{(t+1)}$, then $p(\hat{b}_j^{(t+1)} = 0) \to 1$.

(ii) If $\exists\ c > 0$, s.t. $|\mathrm{corr}(\mathbf{x}_j, \mathbf{y} | \tilde{\mathbf{b}}_{-j}^{(t+1)})| > c$, then $p(\hat{b}_j^{(t+1)} \neq 0) \to 1$.

Iterative Adaptive Lasso (IAL), how to choose δ and τ

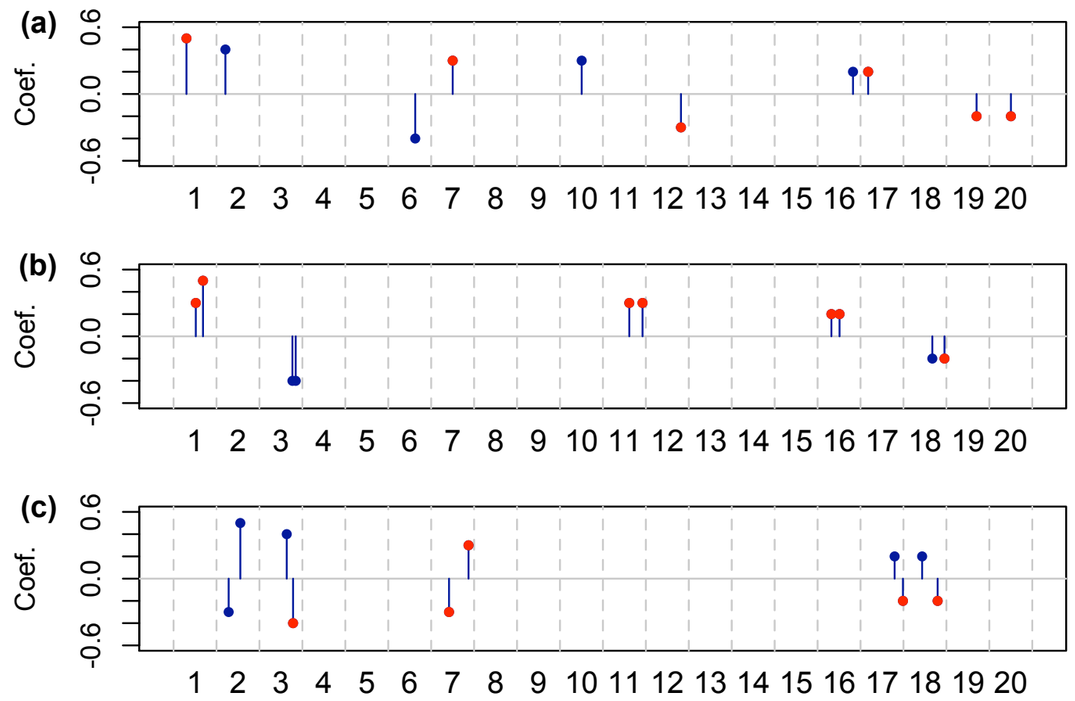
Theorem 1 only provides the magnitude of $(1+\delta)/\tau$. In practice, we selected δ and τ by the following BIC criterion followed by a backward filtering.

$$\mathrm{BIC}_{\tau,\delta} = \log(\mathtt{rss}/n) + \frac{\log(n)}{n} df_{\tau,\delta}$$

where $df_{\tau,\delta}$ is the number of nonzero coefficients.

The backward filtering start from the covariate with the smallest coefficient, compare the model with/without this covariate by ANOVA test. The test p-value can be set as $0.05/p_E$, where p_E is the effective number of independent genotype profiles.

Simulation setup We simulate 20 chromosomes of length 90 cM, with 100 markers per chromosome (using function `sim.map` in R/qtl). Then we simulate genotype data of the 360 F2 mice, and randomly select 1200 markers as markers with "observed genotype profiles".



Results Evaluation: Definition of True Discovery

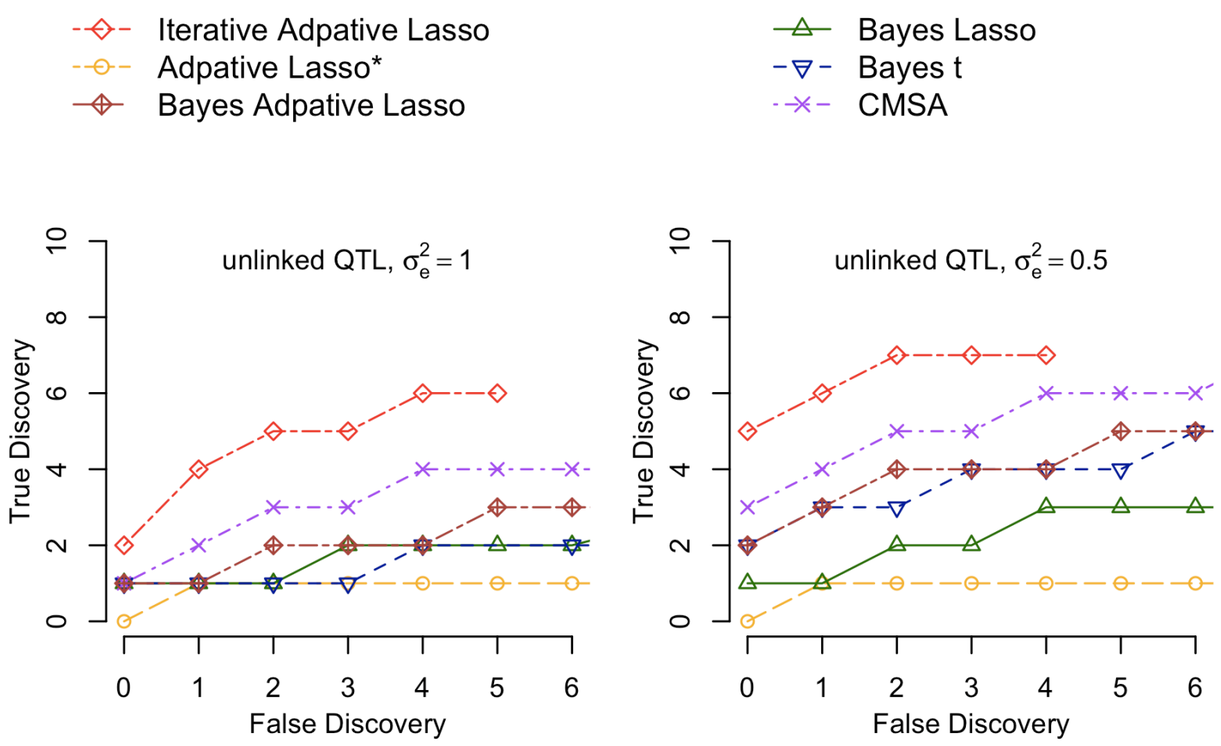
For each of the true QTL, we check whether any SNP in the final model satisfies the following three criteria:

1. it is located on the same chromosome as the QTL
2. it has the same effect direction (sign of the coefficient) as the QTL
3. the r^2 between this SNP and the QTL is bigger than 0.8.

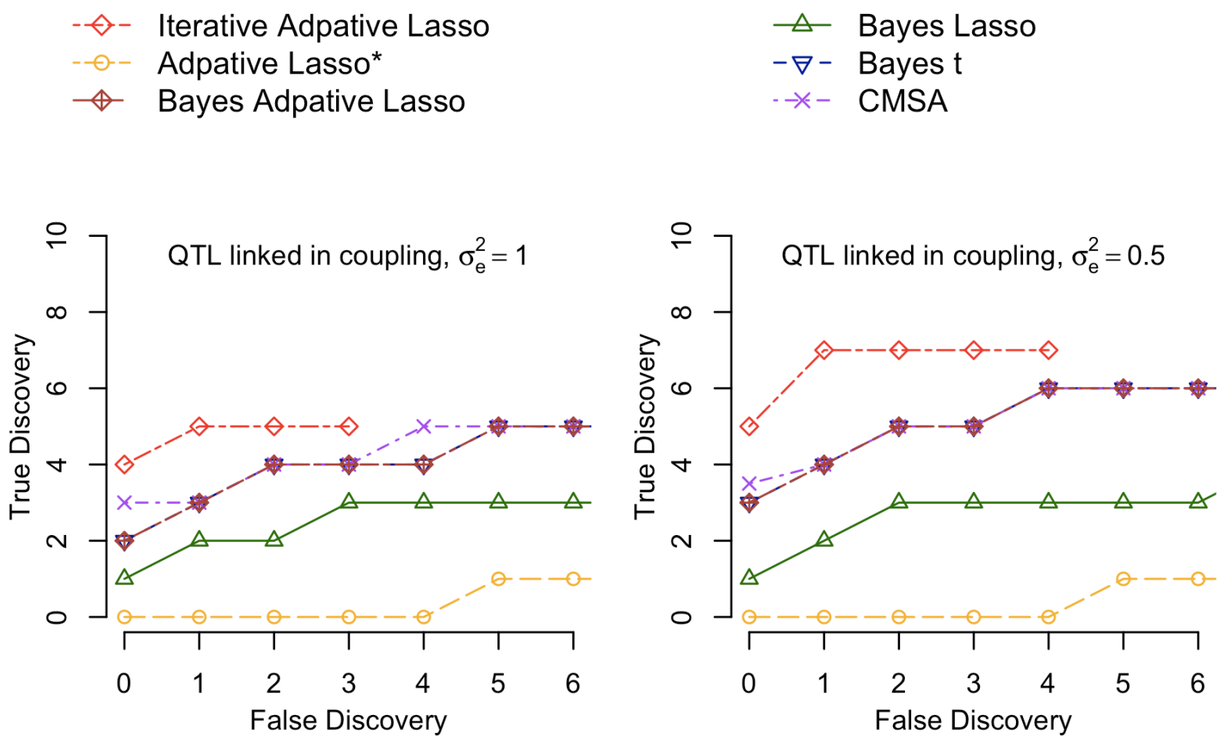
If there is no such SNP, there is no true discovery for this QTL.

If there is at least one such SNP, the SNP with the highest r^2 is recorded as the true discovery of the QTL. The other SNPs are referred to as linked false discoveries.

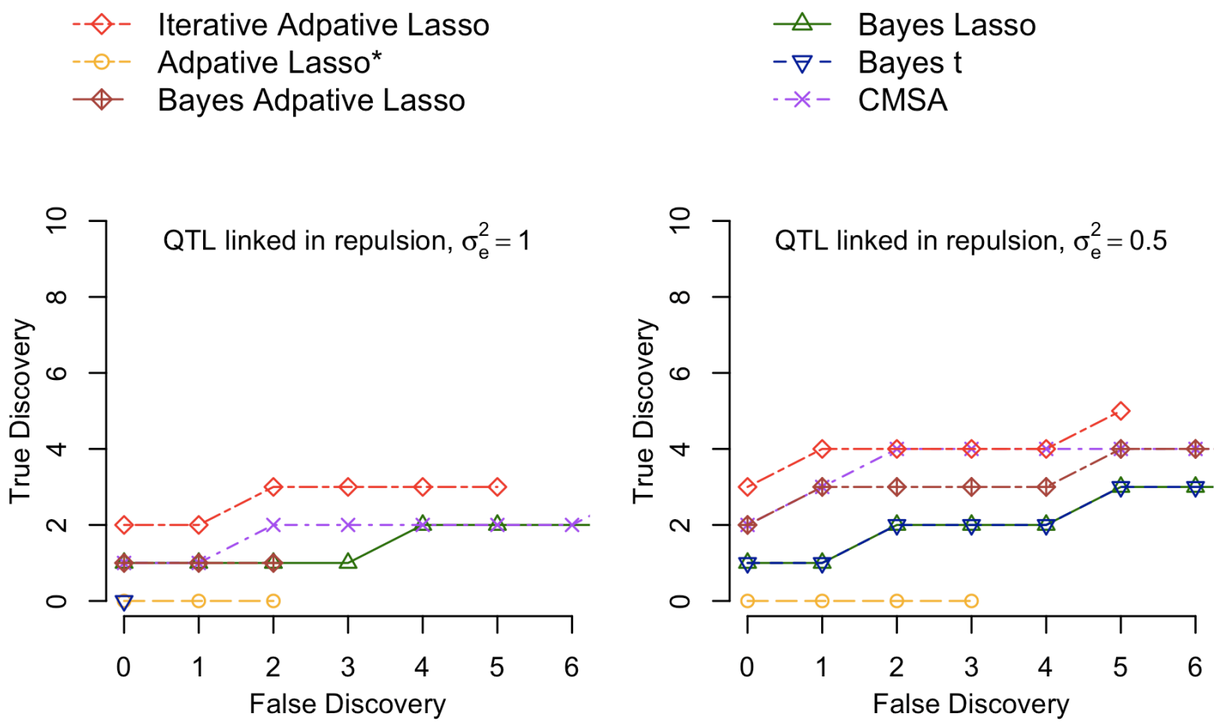
Un-linked QTL



QTL linked in coupling



QTL linked in repulsion



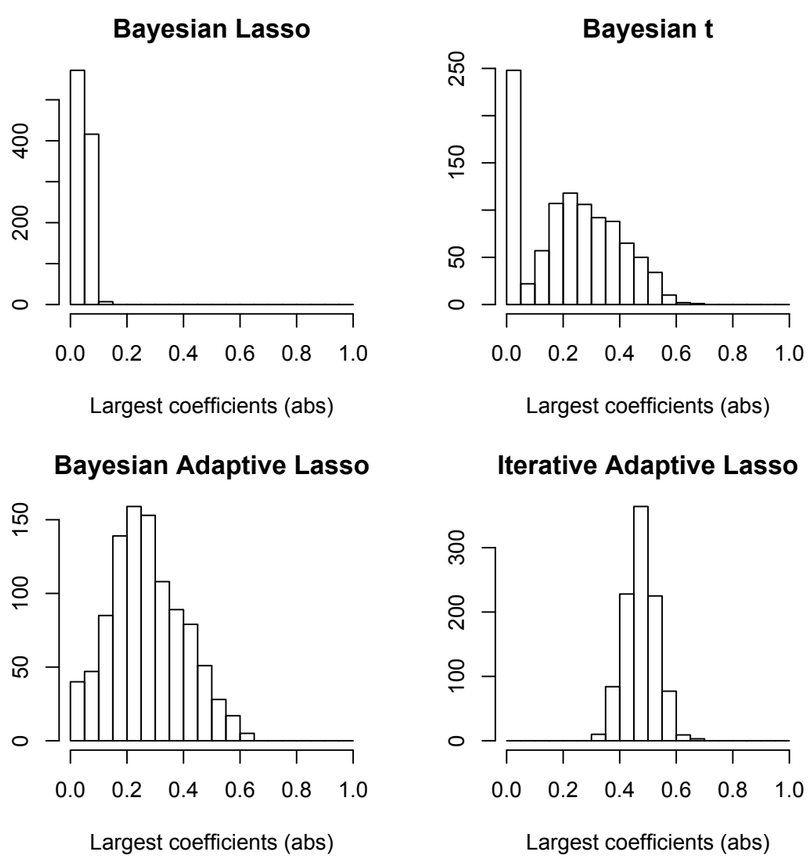
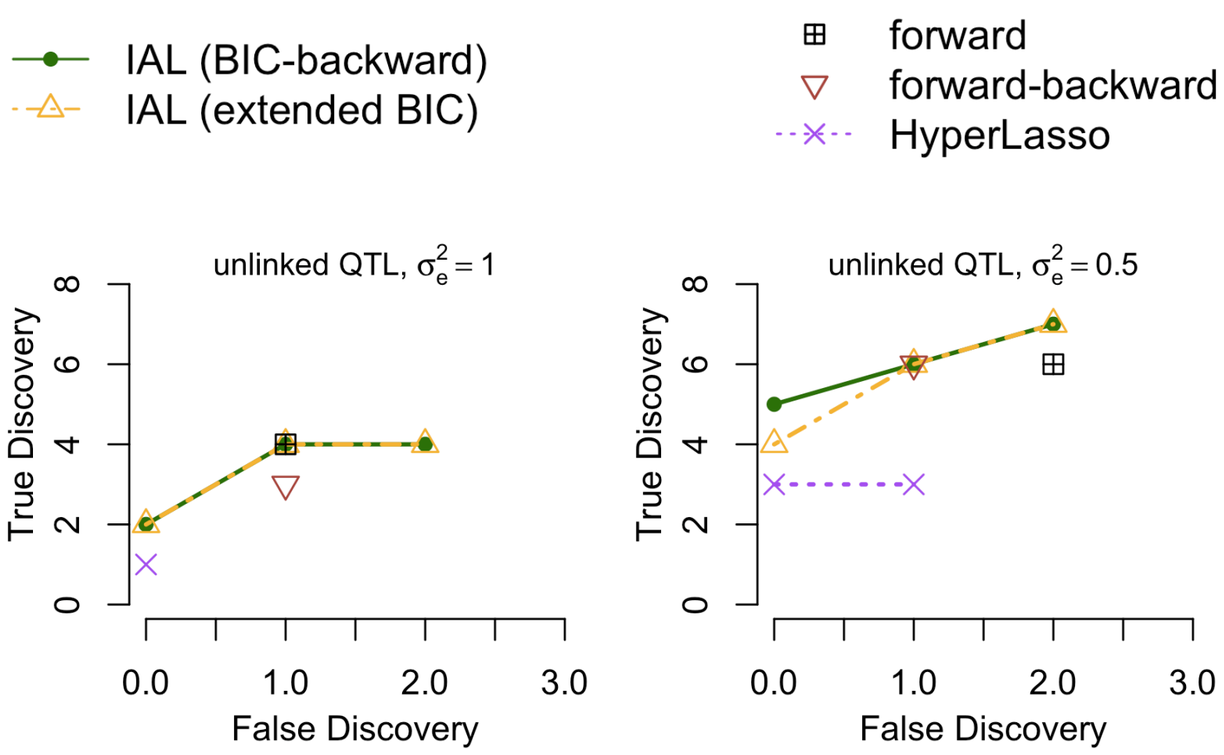
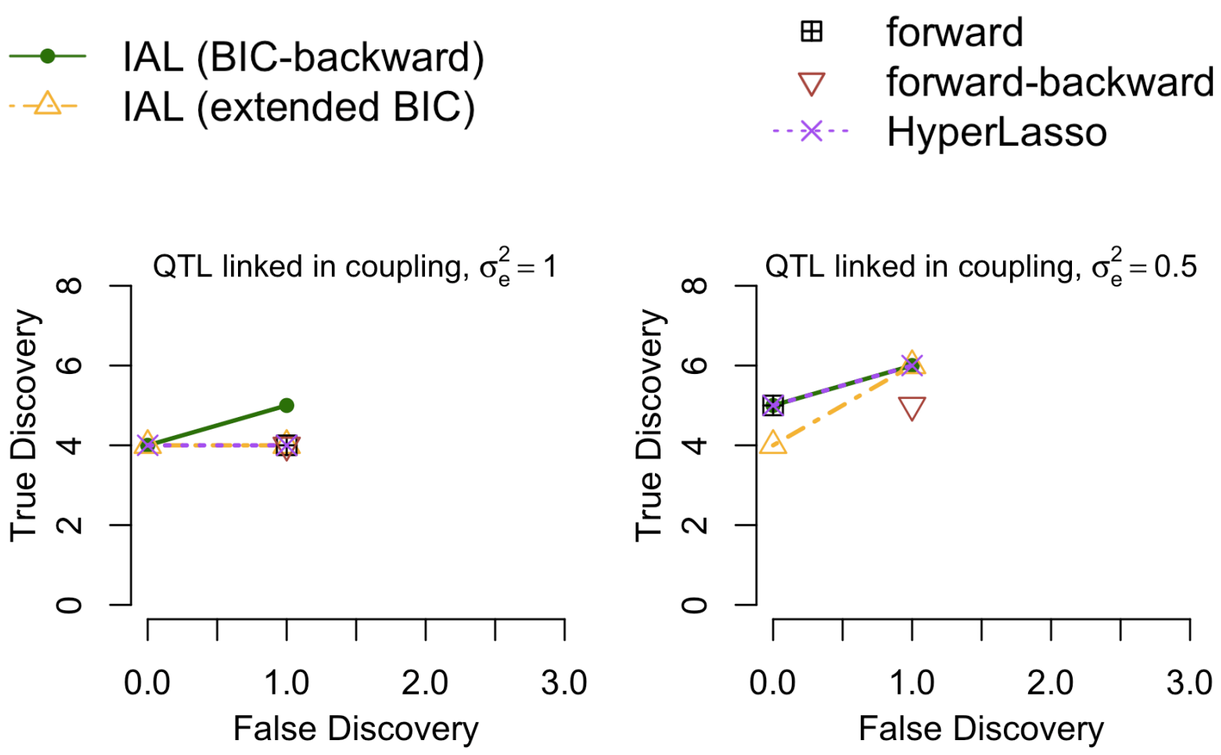


Figure 1: The largest coefficients from 100 simulations in situation 2

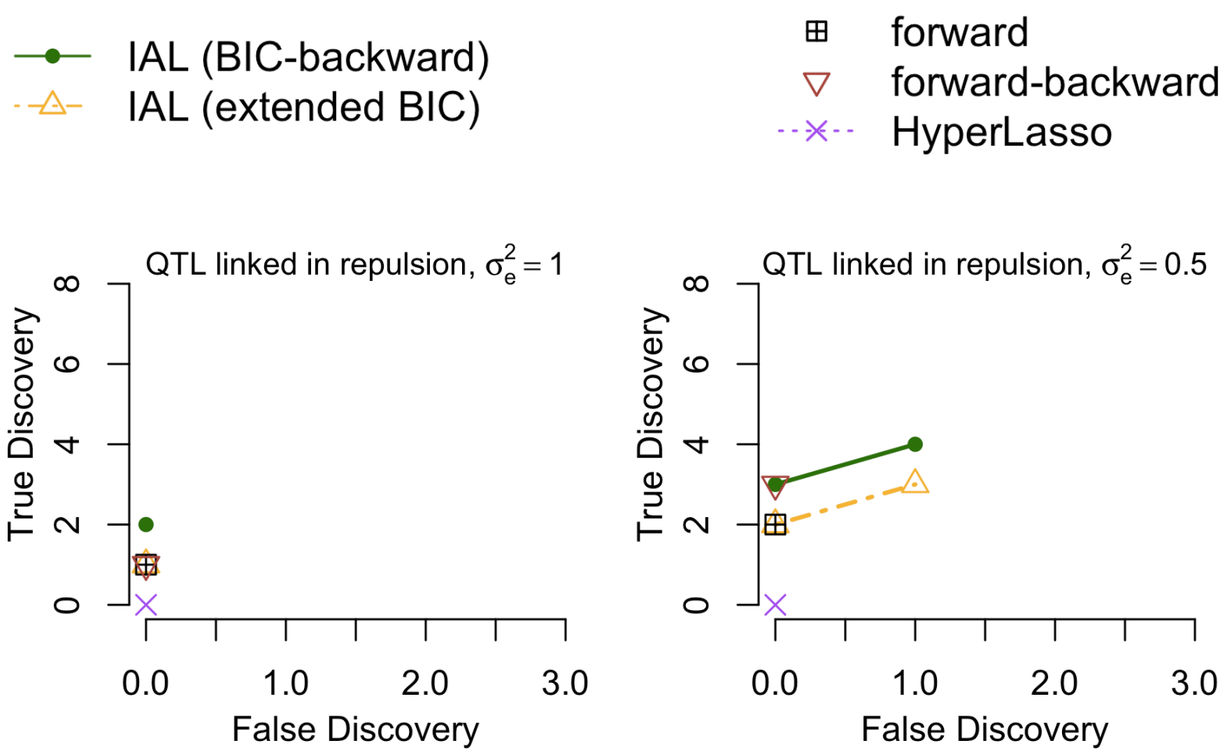
Un-linked QTL



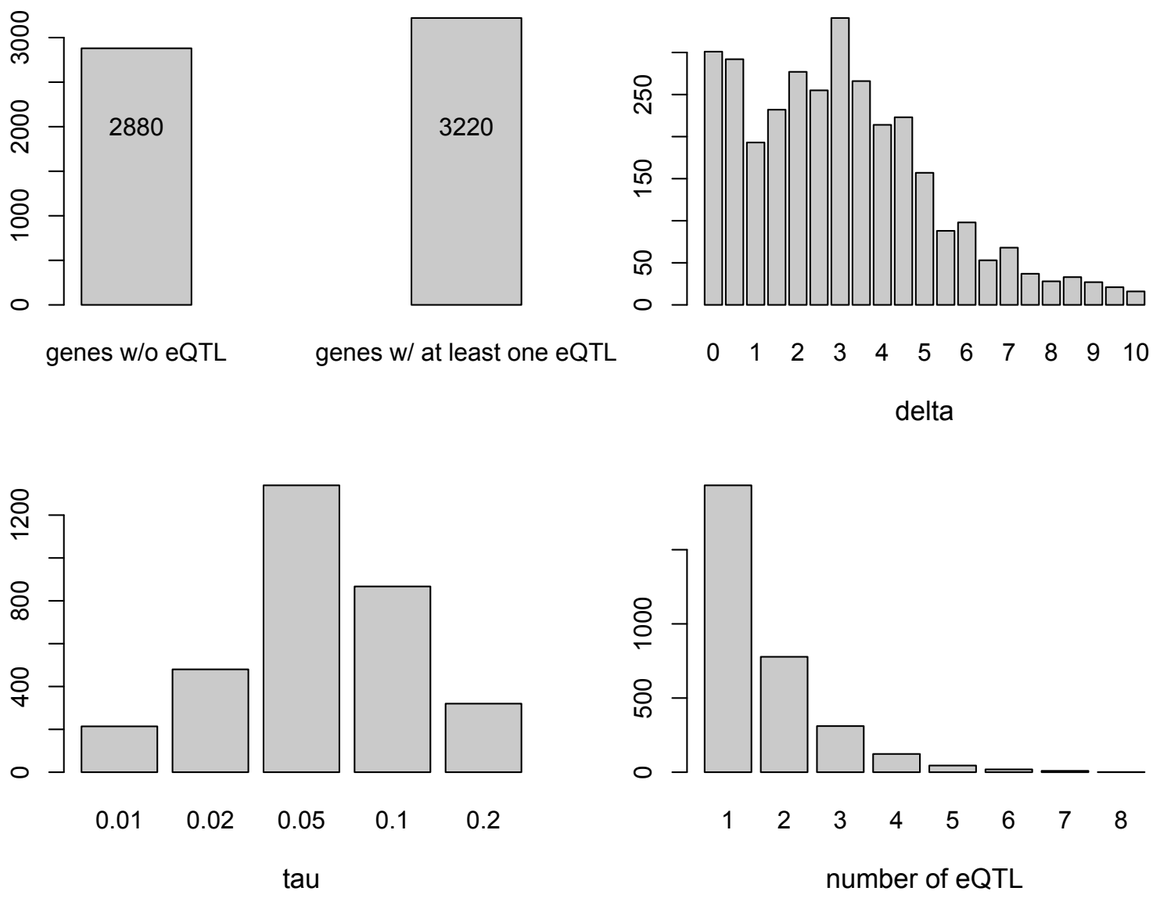
QTL linked in coupling



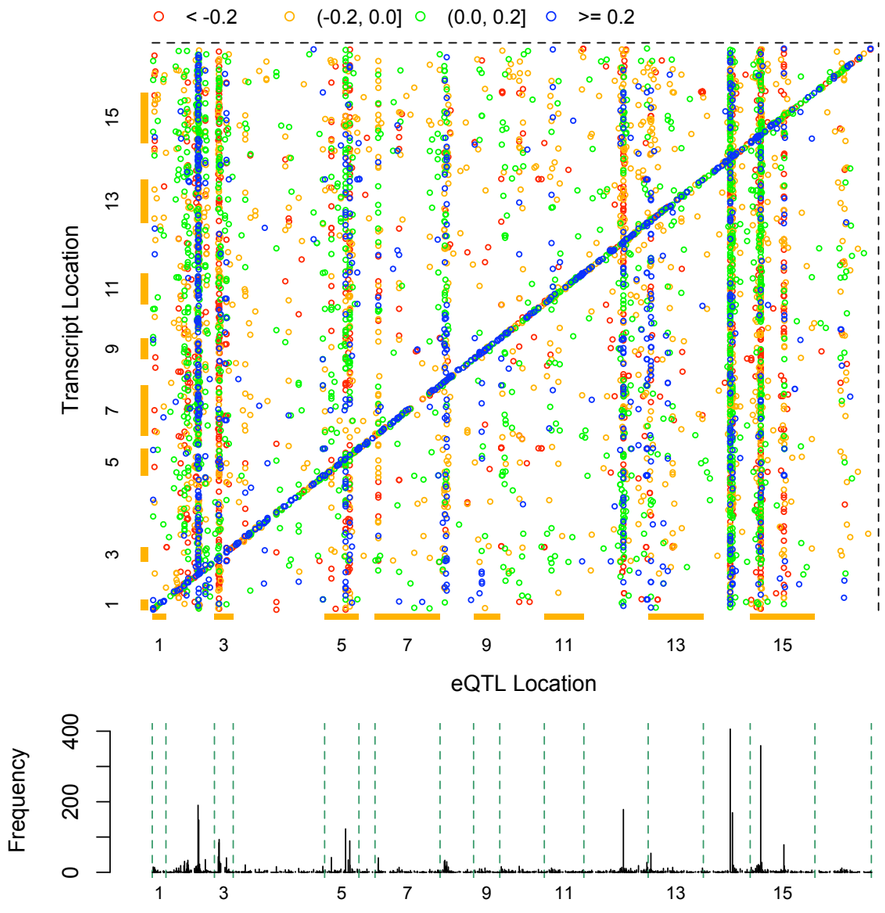
QTL linked in repulsion



Application in eQTL studies



Application in eQTL studies



Application in eQTL studies