SUPPLEMENTARY NOTES
for
A novel variational Bayes multiple locus z-statistic for genome-wide association studies with Bayesian model averaging

Benjamin A. Logsdon[1,*], Cara L. Carty[1], Alexander P. Reiner[12], James Y. Dai[13], Charles Kooperberg[1]
1 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
2 Department of Epidemiology, University of Washington, Seattle, WA, USA
3 Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
∗ E-mail: blogsdon@fhcrc.org

Supplementary Methods

Variational Bayes approximation

For the variational Bayes approximation we wish to estimate the distribution

$$\widehat{Q}\left(\Theta|W\right) = \underset{Q(\Theta|W)}{\operatorname{argmin}} D_{\text{KL}}\left(Q\left(\Theta|W\right)||p\left(\Theta|W\right)\right).$$

Previous authors [1,2] have shown that this minimum is attained when

$$\log\left\{\widehat{q}_j\left(\theta_j|W\right)\right\} = \mathrm{E}_{\widehat{Q}(\Theta_{-j}|W)}\left[\log\left\{p\left(\Theta,W\right)\right\}\right] + C,$$

for all factorized parameters $j = 1, \ldots, J$, where $\mathrm{E}_{\widehat{Q}(\Theta_{-j}|W)}$ represents expectation taken with respect to the estimated approximate posterior distribution removing the j^{th} parameter, and C is a constant to ensure $\int \widehat{q}_j\left(\theta_j|W\right)d\theta_j = 1$. Conveniently, it is possible to estimate $\widehat{Q}\left(\Theta|W\right)$ through coordinate updates of each of these individual approximate distributions until a local minimum of the KL divergence is reached. In addition, for any given local minimum of the KL divergence it is possible to estimate a lower bound on the log marginal probability of the data $\log\left\{p\left(W\right)\right\}$,

$$\mathcal{L}\left(W\right) = \int \widehat{Q}\left(\Theta|W\right)\log\left(\frac{p\left(\Theta,W\right)}{\widehat{Q}\left(\Theta|W\right)}\right)d\Theta.$$

The inequality $\mathcal{L}\left(W\right) \leq \log\left\{p\left(W\right)\right\}$ holds because

$$\begin{aligned}
\log\left\{p\left(W\right)\right\} &= \log\left\{p\left(W\right)\right\} \\
&= \int \widehat{Q}\left(\Theta|W\right)\log\left(\frac{p\left(\Theta,W\right)}{p\left(\Theta|W\right)}\right)d\Theta \\
&= \int \widehat{Q}\left(\Theta|W\right)\log\left(\frac{p\left(\Theta,W\right)\widehat{Q}\left(\Theta|W\right)}{\widehat{Q}\left(\Theta|W\right)p\left(\Theta|W\right)}\right)d\Theta \\
&= \mathcal{L}\left(W\right) + D_{\text{KL}}\left(\widehat{Q}\left(\Theta|W\right)||p\left(\Theta|W\right)\right) \\
&\geq \mathcal{L}\left(W\right),
\end{aligned}$$

since $D_{\text{KL}}\left(\widehat{Q}\left(\Theta|W\right)||p\left(\Theta|W\right)\right) > 0$ except when $\widehat{Q}\left(\Theta|W\right) = p\left(\Theta|W\right)$.

Derivations of vBsr updates

The complete posterior distribution of the vBsr model is

$$p\left(\beta_1,\ldots,\beta_m,\mathbf{y}|\alpha_1,\ldots,\alpha_p,\sigma_e^2,\mathbf{X},\mathbf{Z}\right)=\prod_i^n\frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left\{-\frac{1}{2\sigma_e^2}\left(y_i-\sum_j x_{ij}\beta_j-\sum_k z_{ik}\alpha_k\right)^2\right\}$$

$$\times\prod_j^m p_\beta^{\mathrm{I}[\beta_j\neq 0]}\left(1-p_\beta\right)^{1-\mathrm{I}[\beta_j\neq 0]}.$$

The log-posterior is therefore

$$\log\left\{p\left(\beta_1,\ldots,\beta_m,\mathbf{y}|\alpha_1,\ldots,\alpha_p,\sigma_e^2,\mathbf{X},\mathbf{Z}\right)\right\}=-\frac{n}{2}\log\left\{2\pi\sigma_e^2\right\}-\frac{1}{2\sigma_e^2}\sum_i^n\left(y_i-\sum_j x_{ij}\beta_j-\sum_k z_{ik}\alpha_k\right)^2$$

$$+\log\left(p_\beta\right)\sum_j^m\left(\mathrm{I}\left[\beta_j\neq 0\right]\right)+\log\left(1-p_\beta\right)\sum_j^m\left(1-\mathrm{I}\left[\beta_j\neq 0\right]\right).$$

We can now derive the updates for a given $\widehat{q}_{\beta_j}\left(\beta_j\right)$ approximate posterior distribution,

$$\begin{aligned}\log\left\{\widehat{q}_{\beta_j}\left(\beta_j\right)\right\} &= \mathrm{E}_{\prod_{l\neq j}q_{\beta_l}(\beta_l)}\left[\log\left\{p\left(\beta_1,\ldots,\beta_m,\mathbf{y}|\alpha_1,\ldots,\alpha_p,\sigma_e^2,\mathbf{X},\mathbf{Z}\right)\right\}\right]+C\\ &= -\frac{1}{2\sigma_e^2}\left(\beta_j^2\sum_i^n x_{ij}^2-2\beta_j\sum_i^n x_{ij}\left(y_i-\sum_{l\neq j}x_{il}\mathrm{E}\left[\beta_l\right]-\sum_k z_{ik}\alpha_k\right)\right)\\ &\quad +\log\left(p_\beta\right)\mathrm{I}\left[\beta_j\neq 0\right]+\log\left(1-p_\beta\right)\left(1-\mathrm{I}\left[\beta_j\neq 0\right]\right)+C\\ &= -\frac{1}{2\sigma_j^2}\left(\beta_j-\mu_j\right)^2+\frac{\mu_j^2}{2\sigma_j^2}+\log\left(p_\beta\right)\mathrm{I}\left[\beta_j\neq 0\right]+\log\left(1-p_\beta\right)\left(1-\mathrm{I}\left[\beta_j\neq 0\right]\right)+C,\end{aligned}$$

with $\mu_j=\left(\sum_i x_{ij}^2\right)^{-1}\sum_i x_{ij}\left(y_i-\sum_{k\neq j}x_{ik}p_k^t\mu_k^t-\sum_l z_{il}\alpha_l^t\right)$ and $\sigma_j^2=\left(\sum_i x_{ij}^2\right)^{-1}\sigma_e^2$. This approximate posterior is a mixture of a normal distribution and a point mass with normalization constant

$$\exp\left\{C\right\}=\left(1-p_\beta\right)+p_\beta\sqrt{2\pi\sigma_j^2}\exp\left\{\frac{\mu_j^2}{2\sigma_j^2}\right\}.$$

The approximate posterior mixing probabilities are therefore

$$p_j=\frac{p_\beta\sqrt{2\pi\sigma_j^2}\exp\left\{\frac{\mu_j^2}{2\sigma_j^2}\right\}}{\left(1-p_\beta\right)+p_\beta\sqrt{2\pi\sigma_j^2}\exp\left\{\frac{\mu_j^2}{2\sigma_j^2}\right\}}.$$

We can now specify the lower bound for the vBsr model

$$\mathcal{L}\left(\mathbf{y}|\alpha_1,\ldots,\alpha_p,\sigma_e^2,\mathbf{X},\mathbf{Z}\right)=-\frac{n}{2}\log\left\{2\pi\sigma_e^2\right\}-\frac{1}{2\sigma_e^2}\sum_i^n\mathrm{E}_{\prod_l q_{\beta_l}(\beta_l)}\left[\left(y_i-\sum_j x_{ij}\beta_j-\sum_k z_{ik}\alpha_k\right)^2\right]$$

$$+\sum_j p_j \log\left(\frac{p_\beta\sqrt{2\pi\sigma_j^2}}{p_j}\right)+\sum_j\left(1-p_j\right)\log\left(\frac{1-p_\beta}{1-p_j}\right).$$

The parameter σ_e^2 is estimated through maximization of the lower bound,

$$\begin{aligned}
\frac{d\mathcal{L}}{d\sigma_e^2} &= -\frac{n}{2\sigma_e^2}+\frac{1}{2\sigma_e^4}\sum_i^n \mathrm{E}_{\prod_l q_{\beta_l}(\beta_l)}\left[\left(y_i-\sum_j x_{ij}\beta_j-\sum_k z_{ik}\alpha_k\right)^2\right]\\
0 &= -\frac{n}{2\sigma_e^2}+\frac{1}{2\sigma_e^4}\sum_i^n \mathrm{E}_{\prod_l q_{\beta_l}(\beta_l)}\left[\left(y_i-\sum_j x_{ij}\beta_j-\sum_k z_{ik}\alpha_k\right)^2\right]\\
\sigma_e^2 &= \frac{1}{n}\sum_i^n \mathrm{E}_{\prod_l q_{\beta_l}(\beta_l)}\left[\left(y_i-\sum_j x_{ij}\beta_j-\sum_k z_{ik}\alpha_k\right)^2\right].
\end{aligned}$$

In addition, the parameters $\alpha=\alpha_1,\ldots,\alpha_p$ are also estimated through maximization of the lower bound,

$$\begin{aligned}
\frac{d\mathcal{L}}{d\alpha} &= -\frac{1}{2\sigma_e^2}\left(2\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\alpha-2\mathbf{Z}^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\mathrm{E}_{\prod_{\mathbf{l}}\mathbf{q}_{\beta_{\mathbf{l}}}}[\beta]\right)\right)\\
0 &= -\frac{1}{2\sigma_e^2}\left(2\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\alpha-2\mathbf{Z}^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\mathrm{E}_{\prod_{\mathbf{l}}\mathbf{q}_{\beta_{\mathbf{l}}}}[\beta]\right)\right)\\
\alpha &= \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\mathrm{E}_{\prod_{\mathbf{l}}\mathbf{q}_{\beta_{\mathbf{l}}}}[\beta]\right),
\end{aligned}$$

where $\mathbf{Z}^{\mathrm{T}}$ denotes the transposed matrix representation of the unpenalized variables, and $\mathbf{X}$ denotes the matrix representation of the penalized variables.

vBsr updates

The updates for each regression coefficient β_j for step $t+1$ of the algorithm are

$$\widehat{q_{\beta_j}}\left(\beta_j\right)^{t+1}=\mathrm{I}\left[\beta_j=0\right]\left(1-p_j^{t+1}\right)+\mathrm{I}\left[\beta_j\neq 0\right]p_j^{t+1}\mathcal{N}\left(\mu_j^{t+1},\left(\sigma_j^2\right)^{t+1}\right),$$

where

$$\mu_j^{t+1}=\frac{\sum_i x_{ij}\left(y_i-\sum_{k\neq j}x_{ik}p_k^t\mu_k^t-\sum_l z_{il}\alpha_l^t\right)}{\sum_i x_{ij}^2},$$

$$\left(\sigma_j^2\right)^{t+1}=\frac{\left(\sigma_e^2\right)^t}{\sum_i x_{ij}^2},$$

and

$$p_j^{t+1}=\frac{1}{1+\exp\left\{-\mathcal{G}\left(\ell_0,\left(\sigma_j^2\right)^{t+1},\mu_j^{t+1}\right)\right\}},$$

with

$$\begin{aligned}
\mathcal{G}\left(\ell_0, \left(\sigma_j^2\right)^{t+1}, \mu_j^{t+1}\right) &= \frac{1}{2}\left(\left(\frac{\mu_j^{t+1}}{\sqrt{\left(\sigma_j^2\right)^{t+1}}}\right)^2 + 2\log\left(p_\beta\right) + \log\left(2\pi\left(\sigma_j^2\right)^{t+1}\right) - 2\log\left(1-p_\beta\right)\right) \\
&= \frac{1}{2}\left(\left(\frac{\mu_j^{t+1}}{\sqrt{\left(\sigma_j^2\right)^{t+1}}}\right)^2 + \ell_0 + \log\left(\left(\sigma_j^2\right)^{t+1}\right)\right).
\end{aligned}$$

The expectations for the β_j parameters are

$$\mathrm{E}_{\widehat{q_{\beta_j}}(\beta_j)^{t+1}}\left[\beta_j^t\right] = \langle\beta_j^t\rangle = p_j^t\mu_j^t,$$

and

$$\langle\left(\beta_j^t\right)^2\rangle = p_j^t\left(\left(\mu_j^t\right)^2 + \left(\sigma_j^2\right)^t\right),$$

where $\langle x\rangle$ denotes the expectation of the random variable x with respect to its estimated approximate posterior distribution. Now consider the expected sum of square error term in the approximate log-posterior (where expectations are always taken with respect to the approximate posterior distribution)

$$\begin{aligned}
\langle U^t\rangle &= \sum_i^n \mathrm{E}_{\prod_l \widehat{q_{\beta_l}}(\beta_l)^t}\left[\left(y_i - \sum_j x_{ij}\beta_j - \sum_k z_{ik}\alpha_k\right)^2\right] \\
&= \mathbf{y^T y} - \mathbf{2y^T Z}\left(\alpha^{\mathrm{t}}\right) - \mathbf{2y^T X}\langle\left(\beta^{\mathrm{t}}\right)\rangle - \mathbf{2}\left(\alpha^{\mathrm{t}}\right)^{\mathbf{T}}\mathbf{Z^T X}\langle\left(\beta^{\mathrm{t}}\right)\rangle \\
&\quad + \left(\alpha^{\mathrm{t}}\right)^{\mathbf{T}}\mathbf{Z^T Z}\left(\alpha^{\mathrm{t}}\right) + \langle\left(\beta^{\mathrm{t}}\right)^{\mathbf{T}}\mathbf{X^T X}\left(\beta^{\mathrm{t}}\right)\rangle,
\end{aligned}$$

with $\mathbf{X^T}$ denoting the transposed matrix representation of the predictors. The most difficult term computationally is the last term, which is a product of second moment matrices. Fortunately, because of the mutual factorization among the β_j parameters, we can expand this term as

$$\langle\left(\beta^{\mathrm{t}}\right)^{\mathbf{T}}\mathbf{X^T X}\left(\beta^{\mathrm{t}}\right)\rangle = \langle\left(\beta^{\mathrm{t}}\right)\rangle^{\mathbf{T}}\mathbf{X^T X}\langle\left(\beta^{\mathrm{t}}\right)\rangle - \sum_j\sum_i x_{ij}^2\langle\beta_j^t\rangle^2 + \sum_j\sum_i x_{ij}^2\langle\left(\beta_j^t\right)^2\rangle.$$

With this expected sufficient statistic, we can now define the update for the error variance parameter σ_e^2, by maximizing the lower bound

$$\left(\sigma_e^2\right)^{t+1} = \frac{\langle U^t\rangle}{n}.$$

Finally, the updates for the maximum approximate posterior estimates of the fixed effect parameters, α are

$$\alpha^{t+1} = \left(\mathbf{Z^T Z}\right)^{\mathbf{-1}}\mathbf{Z^T}\left(\mathbf{y} - \mathbf{X}\langle\beta^{\mathrm{t}}\rangle\right).$$

As this optimization problem is not convex, it is not guaranteed to converge to a global maximum. In practice, we start the algorithm using starting values for all the β_j sufficient statistics (μ_j, σ_j, p_j) initialized at the origin, σ_e^2 initialized to the variance of the observed response, and α_k initialized to the origin. We continue the algorithm until convergence is assessed (as described below). To deal with the non-convexity, we use a number of random re-orderings of the SNPs and average the results. Details of this are given below.

Derivation of the z_{vb} statistic

Consider the limit of the lower bound $\mathcal{L}$ when the penalty parameter $p_\beta \to 0$ ($\ell_0 \to -\infty$),

$$\begin{aligned}\lim_{p_\beta \to 0} \mathcal{L} &\to -\frac{n}{2}\log\left\{2\pi\sigma_e^2\right\} - \frac{1}{2\sigma_e^2}\sum_i^n \left(y_i - \sum_k z_{ik}\alpha_k\right)^2 \\ &\to \ell\left(\beta_1 = 0, \ldots, \beta_m = 0, \sigma_e^2, \alpha\right) \\ &\to \ell\left(\beta_j = 0, \sigma_e^2, \alpha\right),\end{aligned}$$

since $\langle U \rangle \to \sum_i^n \left(y_i - \sum_k z_{ik}\alpha_k\right)^2$, $\sum_j p_j \log\left(\frac{p_\beta\sqrt{2\pi\sigma_j^2}}{p_j}\right) \to 0$, $\sum_j (1-p_j)\log\left(\frac{1-p_\beta}{1-p_j}\right) \to 0$, with the log-likelihood function of an unpenalized linear model, $\ell\left(\beta = 0, \sigma_e^2, \alpha\right)$. In this limit of the model with no active penalized variables, the lower bound becomes the unpenalized log-likelihood function associated with a linear model for the unpenalized covariates. If we now consider the score and observed information of this log-likelihood function for the j^{th} penalized variable,

$$U\left(\beta_j\right) = \frac{d\ell\left(\beta_j\right)}{d\beta_j} = -\frac{1}{2\sigma_e^2}\left(2\beta_j\sum_i^n x_{ij}^2 - 2\sum_i\left(y_i - \sum_k z_{ik}\alpha_k\right)\right),$$

$$I\left(\beta_j\right) = -\frac{d^2\ell\left(\beta_j\right)}{d\beta_j^2} = \frac{\sum_i^n x_{ij}^2}{\sigma_e^2},$$

we can define a score statistic for the test of marginal association between the j^{th} variable and the phenotype,

$$S\left(\beta_j = 0\right) = \frac{\sum_i^n x_{ij}\left(y_i - \sum_k z_{ik}\alpha_k\right)}{\sqrt{\sigma_e^2\sum_i^n x_{ij}^2}}.$$

In addition, when $p_\beta \to 0$, the mean and variance updates of the approximate posterior distribution of β_j can be rewritten

$$\mu_j^{t+1} = \frac{\sum_i x_{ij}\left(y_i - \sum_l z_{il}\alpha_l^t\right)}{\sum_i x_{ij}^2},$$

and

$$\left(\sigma_j^2\right)^{t+1} = \frac{\left(\sigma_e^2\right)^t}{\sum_i x_{ij}^2},$$

since $\lim_{p_\beta \to 0} p_k^{t+1} \to 0$ for all markers for a given data-set. Since α and σ_e^2 are estimated through maximization of the lower bound, in this limit they become their respective standard maximum likelihood estimates. We now define the following statistic based on the updates of the approximate posterior distribution,

$$z_j = \frac{\mu_j}{\sqrt{\sigma_j^2}}. \tag{1}$$

When $p_\beta \to 0$ this statistic is equivalent to the score statistic from a marginal analysis of the data and will by asymptotically $\mathcal{N}(0,1)$ by standard arguments,

$$\lim_{p_\beta \to 0} z_j = \frac{\left(\sum_i x_{ij}\left(y_i - \sum_l z_{il}\alpha_l\right)\right)}{\sqrt{\sigma_e^2\sum_i^n x_{ij}^2}} = S\left(\beta_j = 0\right).$$

We now consider the case when the penalty parameter p_β increases from zero. We do not claim that in this case the asymptotic distribution of z_j under the null is always $\mathcal{N}(0,1)$. Yet, we do claim that if certain conditions concerning the data are true, there exists a value of $p_\beta > 0$ ($\ell_0 > -\infty$) where the distribution of most individual statistics in the data is indistinguishable from a $\mathcal{N}(0,1)$. These conditions first include the assumption that a majority of the penalized variables are irrelevant and will therefore have individual statistics (z_j) that will be distributed under the model of no association. Second, we assume that as p_β increases the z_j distribution of the null features does not change drastically from a $\mathcal{N}(0,1)$ until the model becomes over-fit. If these assumptions are true, we can use the empirical cumulative density function (ECDF) of the test statistic defined in Equation (1) to ensure that as more genetic markers enter the model with larger p_k^{t+1} parameters for increasing p_β, the asymptotic distributional assumption of the test-statistic for the null features in the data-set is not violated. This assumes that if too many features are included in the model (so that it is over-fit), then the distribution of this test statistic for the majority of genetic markers will deviate drastically from a $\mathcal{N}(0,1)$ distribution as the effective degrees of freedom of the test increases. Because the ECDF converges uniformly at rate $\sqrt{n}$, we argue that this procedure will be very sensitive to model over-fitting by including too many features in the model, given the assumptions concerning the data are true.

Bayesian model averaging

The lower bound specified by minimizing the KL-divergence, which is computed every iteration through all the parameters, is

$$\mathcal{L}^{t+1} = -\frac{n}{2}\left(\log\left\{2\pi\left(\sigma_e^2\right)^{t+1}\right\}+1\right)+\sum_j p_j^{t+1}\log\left(\frac{p_\beta\sqrt{2\pi\sigma_j^2}}{p_j^{t+1}}\right)+\sum_j\left(1-p_j^{t+1}\right)\log\left(\frac{1-p_\beta}{1-p_j^{t+1}}\right).$$

We stop updating the distributions once $|\mathcal{L}^{t+1}-\mathcal{L}^t| < 10^{-4}$. In addition, after running the algorithm with many different starting orderings, and identifying a set of multiple unique modes in the approximate posterior surface, we perform approximate Bayesian model averaging, by computing the posterior probability for each of the models represented by a unique mode as

$$p\left(M_s\right) = \frac{\exp\left(\mathcal{L}_s\right)}{\sum_s \exp\left(\mathcal{L}_s\right)}.$$

Based on this estimate of the posterior probability of each unique model identified, we re-weight all parameters to produce the final vBsr estimate for each of the genotypes in the model,

$$\begin{aligned}\widehat{\mu_j} &= \sum_s^g p\left(M_s\right)\mu_{sj}\\ \widehat{\sigma_j^2} &= \sum_s^g p\left(M_s\right)\sigma_{sj}^2\\ \widehat{p_j} &= \sum_s^g p\left(M_s\right)p_{sj},\end{aligned}$$

where $\mu_{sj}, \sigma_{sj}^2, p_{sj}$ are the respective sufficient statistics μ_j^t, $\left(\sigma_j^2\right)^t, p_j^t$ at convergence from the s^{th} unique mode identified, of g total unique modes. The Bayesian model averaged $\widehat{z_{vb}}$ statistic for each genotype is defined as

$$\widehat{z_{vb}} = \sum_s^g p\left(M_s\right)\left(\frac{\mu_{sj}}{\sqrt{\sigma_{sj}^2}}\right).$$

If there is a large quantity of model uncertainty, the model variance of the $\widehat{z_{vb}}$ test statistic under the null model will be less than one, depending on the covariance between the estimates across modes of the approximate posterior distribution,

$$\mathrm{Var}\left(\widehat{z_{vb}}\right) = \sum_{s}^{g} p\left(M_s\right)^2 + 2\sum_{k=1}^{g-1}\sum_{l=k+1}^{g} p\left(M_k\right) p\left(M_l\right) \mathrm{Cov}\left(\frac{\mu_{kj}}{\sqrt{\sigma^2_{kj}}}, \frac{\mu_{lj}}{\sqrt{\sigma^2_{lj}}}\right).$$

For the features with effects that are well approximated by the null distribution, their relative contribution to the fit of the model will be very low (since $\langle\beta_j\rangle \approx 0$), and therefore the correlation between their test statistics will be high, and the variance will be close to one. For the results presented in this paper we do not correct for the reduction in variance of the test statistic due to the Bayesian model averaging, though in principal it can be performed by estimating the correlation between the model residuals without the expected effect of a given feature. The approximation we use by not correcting for the reduction in variance will slightly reduce the power of our approach, and overall make it moderately more conservative.

Choice of model complexity parameter ℓ_0

For all simulations and data analyses we solve the vBsr penalized regression model along a path of the model complexity parameter ℓ_0. A marginal analysis is performed initially (i.e. $\langle\beta_k\rangle = 0, \forall k$), and the most extreme value of the following marginal test statistic is used as the starting value of the path

$$-\ell_0 = \max\left(\frac{\mu_j^2}{\sigma_j} + \log\left(\sigma_j\right)\right).$$

The end point of the path is similarly chosen based on the $\sqrt{n}$ largest value of this marginal statistic if $\sqrt{n} < m$, and as the smallest value of the marginal statistic otherwise, for the simulations with independent genotypes. For the simulations and data analysis with correlated genotypes we use a path of $\ell_0 = (-50, \ldots, -8)$ and $\ell_0 = (-30, \ldots, -8)$ respectively, since the $\sqrt{n}$ largest marginal statistic for correlated genotypes was not large enough to identify the best possible models. We use a path length of 50 for simulations and data analyses. For the simulations, where we did not use any marginal pre-screening before running our analysis, we computed the Kullback-Leibler divergence between the observed and expected distribution of $\widehat{z_{vb}}$ up to the 99^{th} percentile as

$$\mathrm{KL} = \frac{\widehat{\omega}^2}{2} + \frac{1}{2}\left(\frac{\widehat{\zeta}^2}{\sigma^2_{ref}} - 1 - \log\left(\frac{\widehat{\zeta}}{\sigma^2_{ref}}\right)\right),$$

where $\hat{\omega} = \mathrm{E}\left(\widehat{z_{vb}} | \mathrm{abs}\left(\widehat{z_{vb}}\right) < c\right)$, $\hat{\zeta} = \mathrm{Var}\left(\widehat{z_{vb}} | \mathrm{abs}\left(\widehat{z_{vb}}\right) < c\right)$, $\sigma^2_{ref} = 1 - \frac{2c\phi(c)}{\Phi(c) - \Phi(-c)}$, and $c = \Phi^{-1}(0.995)$. This statistic was used to capture the deviation of the distribution of $\widehat{z_{vb}}$ for genetic variables that are most likely null from its neutral expectation. While this is not the only goodness of fit measure that could be used, based on our simulations it appeared sufficient in terms of determining when the model starts to become over-fit. To test the flexibility of this statistic, we investigated alternative strategies for choosing this statistic based on its behavior along the path of ℓ_0. Besides choosing just based on the minimum (which appeared to be the most conservative strategy), we investigated the behavior of this statistic under the null model through Monte Carlo simulations. Based on these Monte Carlo simulations we devised a set of both conservative and liberal strategies to choose the statistic that capture the expected behavior of KL under the null. The conservative strategies attempt to find the value of ℓ_0 that most closely match the null expectation, and liberal strategies find ℓ_0 that match the null expectation plus a factor related to the variability in KL. These six different strategies for choosing which value of KL to use along the ℓ_0 path included choosing ℓ_0 with the minimum KL along the path (z_{vb}^{a}), choosing the largest ℓ_0 with $\log(\mathrm{KL})$ less than the expected value of $\mathrm{E}\left(\log(\mathrm{KL})\right)$ (z_{vb}^{b}), the largest value of ℓ_0 with $\log(\mathrm{KL})$ less than $\min\left(\log(\mathrm{KL})\right) + \sqrt{\mathrm{Var}\left(\log(\mathrm{KL})\right)}$ (z_{vb}^{c}), the largest value of ℓ_0 with $\log(\mathrm{KL})$ less than $\mathrm{E}\left(\log(\mathrm{KL})\right) + \sqrt{\mathrm{Var}\left(\log(\mathrm{KL})\right)}$ (z_{vb}^{d}), the largest value of ℓ_0 with $\log(\mathrm{KL})$ less than $\min\left(\log(\mathrm{KL})\right) + 2\sqrt{\mathrm{Var}\left(\log(\mathrm{KL})\right)}$ (z_{vb}^{e}), and the largest value of ℓ_0 with $\log(\mathrm{KL})$ less than $\mathrm{E}\left(\log(\mathrm{KL})\right)+2\sqrt{\mathrm{Var}\left(\log(\mathrm{KL})\right)}$ (z_{vb}^{f}). The expectation and variance of $\log(\mathrm{KL})$ under the null model were computed based on 1000 Monte Carlo simulations. For the data analysis, where the marginal pre-screening was performed by only choosing genetic markers with $P_{sma} \leq 10^{-3}$, we compute the KL-divergence statistic based on the fit of the distribution between the 99.9^{th} and 99.99^{th} percentile of the distribution. Therefore, in this case $\hat{\omega} = \mathrm{E}\left(\widehat{z_{vb}}|\mathrm{abs}\left(\widehat{z_{vb}}\right) < c, \mathrm{abs}\left(\widehat{z_{vb}}\right) > b\right)$, $\hat{\zeta} = \mathrm{Var}\left(\widehat{z_{vb}}|\mathrm{abs}\left(\widehat{z_{vb}}\right) < c, \mathrm{abs}\left(\widehat{z_{vb}}\right) > b\right)$, $\sigma_{ref}^2 = 1 + \frac{c\phi(c)-b\phi(b)}{\Phi(b)-\Phi(c)}$, $b = \Phi^{-1}(0.9995)$, and $c = \Phi^{-1}(0.99995)$.

Supplementary Results

We also investigated a case where the independence among markers assumption was relaxed, by using non overlapping sub-sets of the WHI SHARe genotype data for the first 10^4 markers with minor-allele frequency greater than 0.05 (that passed all the quality control filters) on chromosome 1. We simulated 100 replicate genetic architectures for different sample sizes and total heritabilities as in our other simulations (i.e. 50 causal loci, sample sizes of either 500, 1000, or 2000, fixed heritability of 0.5 or 0.9), but we only considered the vBsr and single-marker analysis approaches, because of the observed sensitivity of the lasso to correlations between true and false positives. We illustrate the performance of the z_{vb} and the χ_{sma}^2 statistic in terms of the FWER and power in Figures S2 and S3 (for the liberal choice of model size, z_{vb}^{f}). In these figures we define a positive when it has a correlation larger than R^2 with the causal locus, for different values of R^2. For both the vBsr statistic and the single-marker analysis approach we set the significance cutoff to control the FWER to 0.05. In general we see in Figure S2 that the vBsr approach can control the FWER much better than single-marker analysis, across a range of R^2 cutoffs, indicating it removes some spurious associations because of linkage disequilibrium. We also see in Figure S3 that in the case of highly heritable phenotypes (i.e. $h^2 = 0.9$), it also has greater power than the single-marker analysis approach.

Because our primary goal is to control the type I error, we wanted to be sure that our choice of the penalty parameter ℓ_0 was appropriate based on the empirical distribution of the test statistic. We show the Quantile-Quantile plots of null features averaged across the 1000 replicate $h^2 = 0.9, n = 500$ simulations for three different strategies for choosing model size (z_{vb}^{a},z_{vb}^{b}, and z_{vb}^{f}) in Figure S4. We see in the left panels of this figure, that the Q-Q plot restricted to the null genotypes corresponds very well to the null distribution assumption of the z_{vb} statistic for the more conservative strategies, and only starts to deviate slightly for the more liberal strategy (top right panel of Figure S4). For illustration purposes, we show these Q-Q plots for the single-marker analysis results averaged across the same 1000 replicated $h^2 = 0.9, n = 500$ simulations for the same null markers in the bottom right panel of Figure S4.

References

[1] Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University of London.

[2] Bishop, C. (2006). *Pattern recognition and machine learning*. Springer, New York.

Supplementary Figures

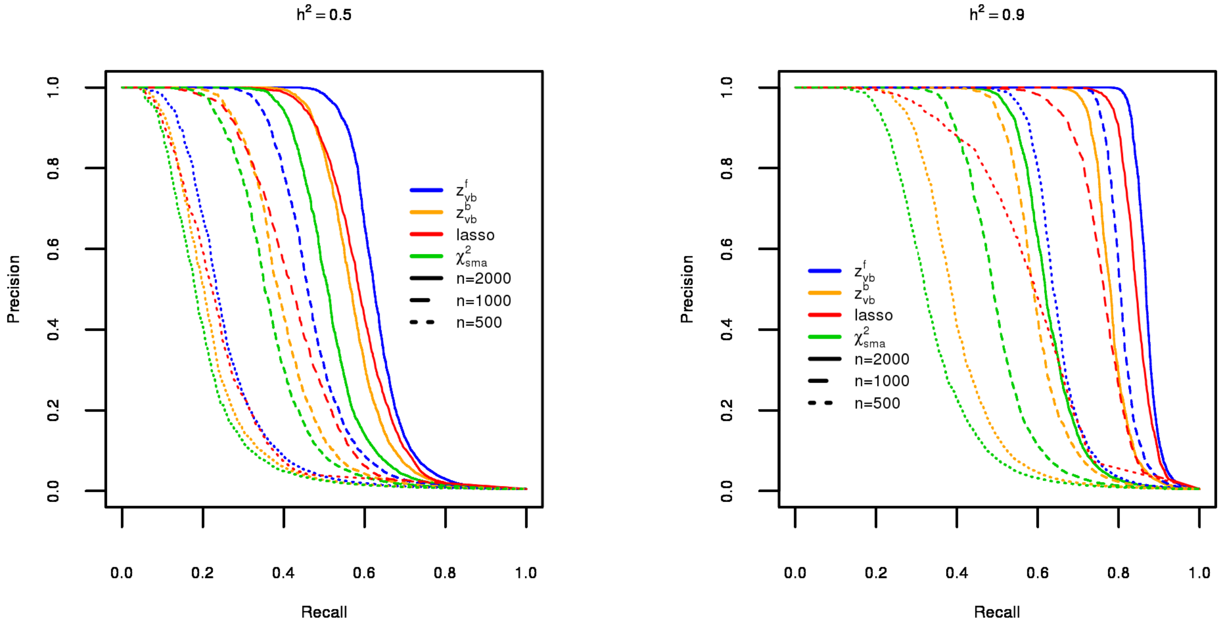

Figure S1: Precision-recall curves are illustrated for the 100 replicate simulations, with 10^4 independent markers, and 50 causal loci for all three methods (using the liberal strategy for choosing model size, based on the expectation of the diagnostic statistic plus two standard errors, z_{vb}^f). Precision is defined as $\frac{tp}{tp+fp}$, and Recall is defined as $\frac{tp}{tp+fn}$, with tp being the number of true positives, fp being the number false positives, and fn being the number of false negatives.

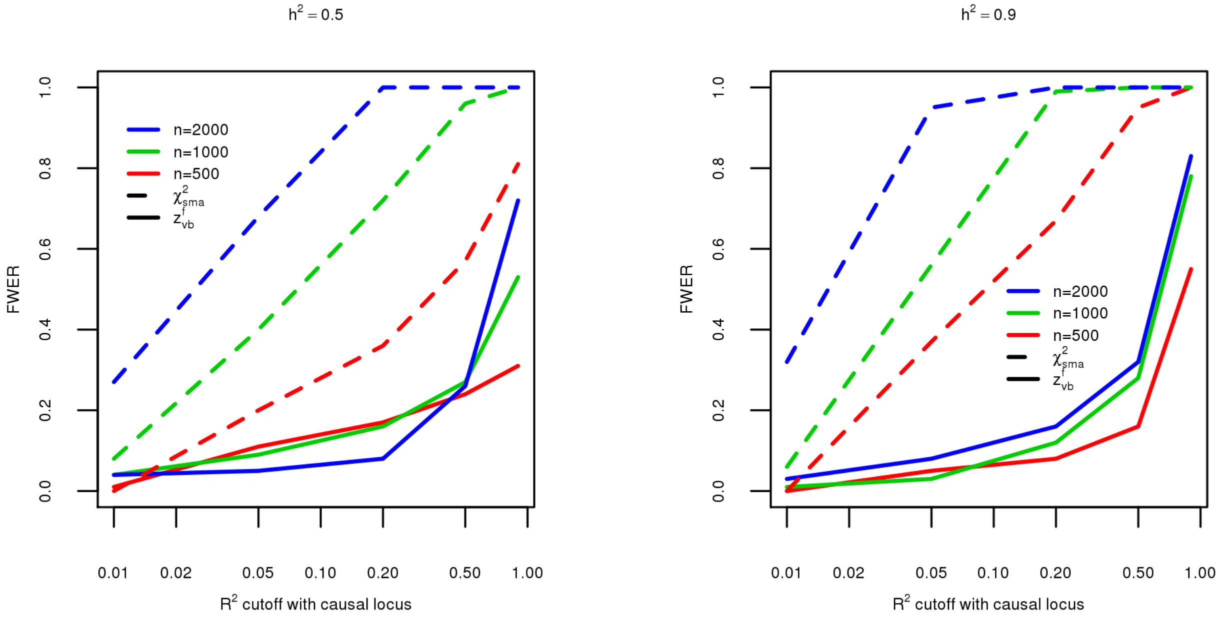

Figure S2: The family-wise error-rate (FWER) for different sample sizes, and fixed total heritabilities are illustrated for different cut-offs of the pairwise R^2 between the causal loci and tagging loci. Results are shown averaged over 100 replicates, with 50 causal loci, with genetic effects sampled from a standard normal distribution.

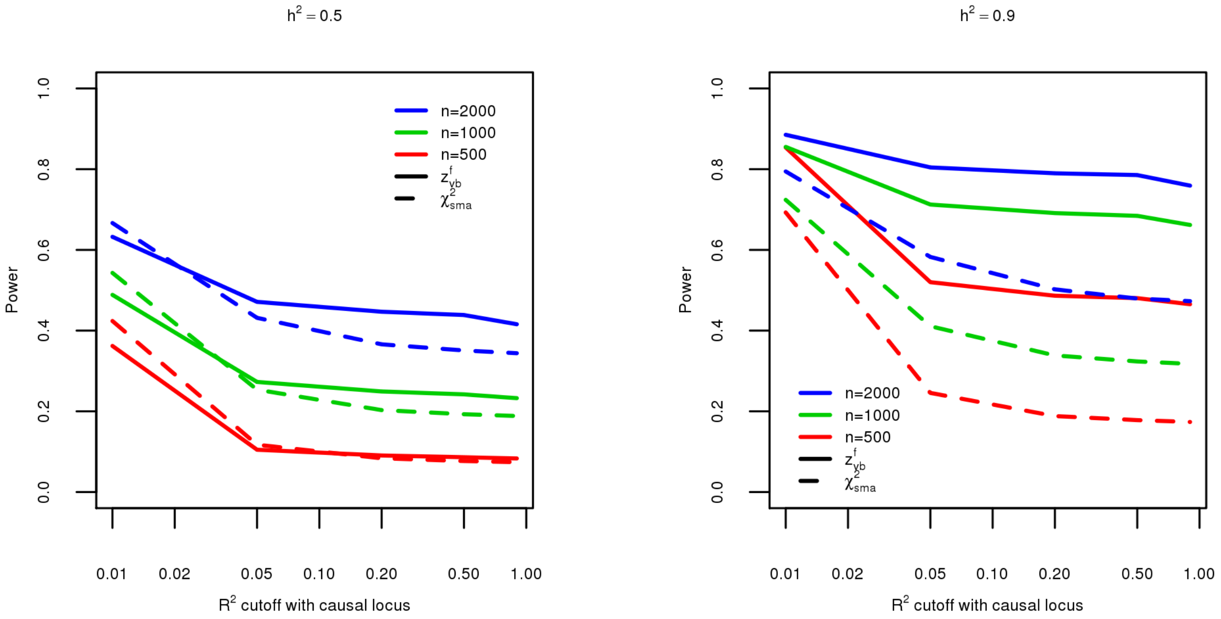

Figure S3: The power for different sample sizes, and fixed total heritabilities are illustrated for different cut-offs of the pairwise R^2 between the causal loci and tagging loci. Results are shown averaged over 100 replicates, with 50 causal loci, with genetic effects sampled from a standard normal distribution.

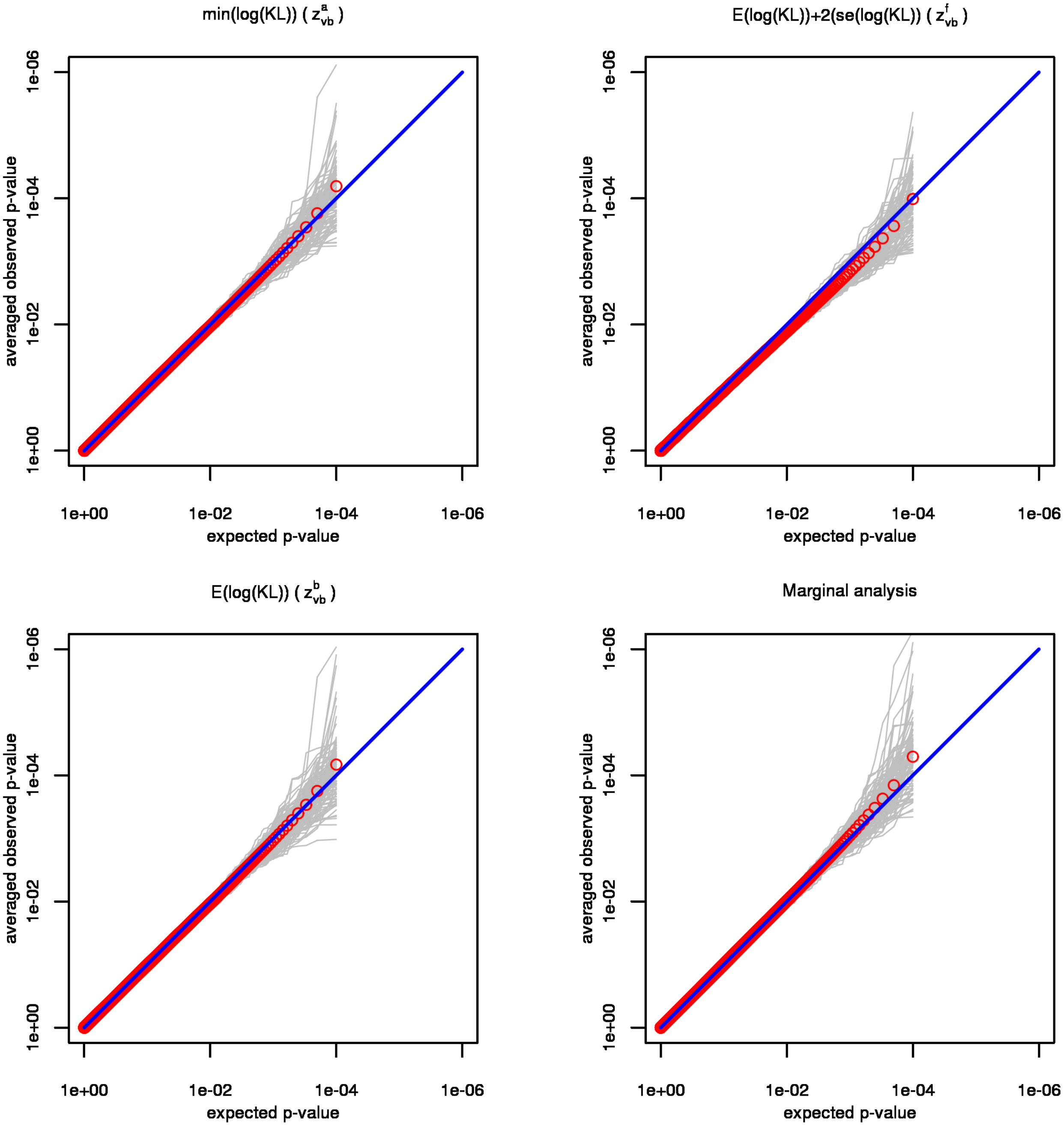

Figure S4: The expected quantiles plotted against the average observed quantiles across 1000 replicates for three different strategies for choosing the model size of the z_{vb} statistic for a simulation with 10^4 independent genotypes, heritability h^2 of 0.9, and sample size $n = 500$. The left panels shows the null distribution of the test statistic for z_{vb} across the null markers for the more conservative strategies, the top right panel shows a more liberal strategy, and the bottom right panel shows the distribution of the test statistic for χ^2_{sma} across the null markers. In light gray are the Q-Q plots of the first 100 replicates for each statistic.

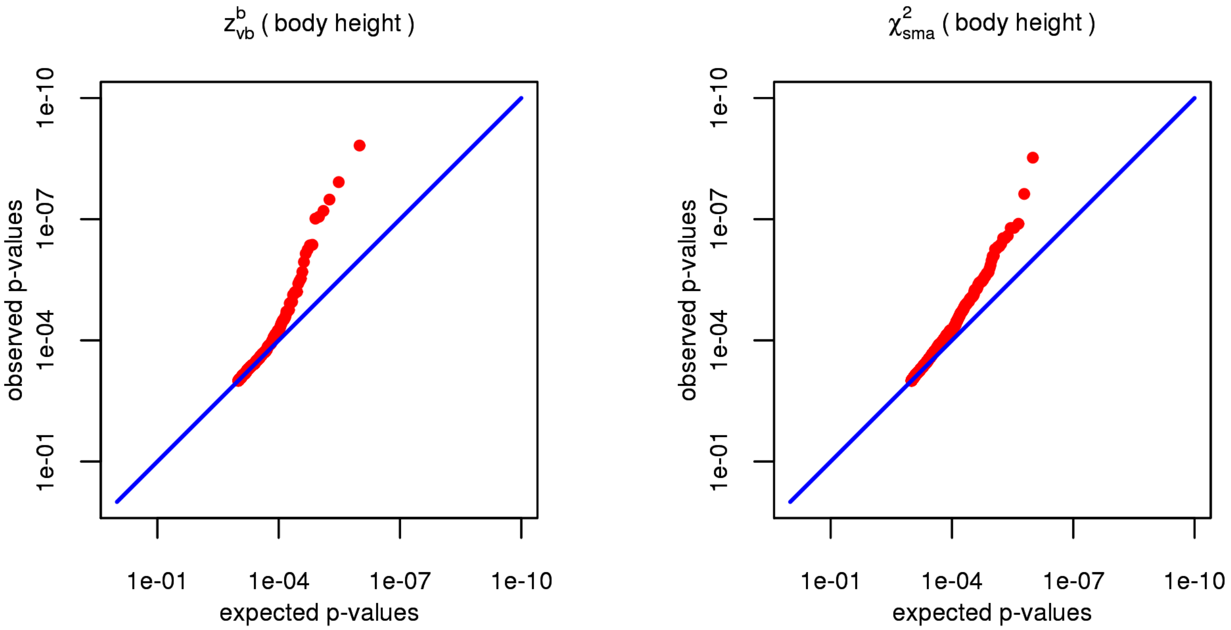

Figure S5: The left panel illustrates the Q-Q plot for the z_{vb} statistic, and the right panel illustrates the Q-Q plot for χ^2_{sma} statistic, both truncated at $P_{vb} < 10^{-3}$ and $P_{sma} < 10^{-3}$ respectively. The vBsr analysis was run with only the genotypes that passed a marginal test $P_{sma} < 10^{-3}$.