note0902

Daniel Sorensen

July 25, 2024

Example. Rare Diseases: Gene Discovery and Prediction

An important topic of modern genomic studies is prediction of a disease before its onset to help with early diagnosis and prevention. This example illustrates some of the issues that may arise in such studies. A setup could be as follows. An analyst has access to randomly drawn data from a homogeneous population of nominally unrelated individuals, recorded for the presence/absence of a particular disease; individuals are genotyped for a large number of genetic markers. An estimate of heritability on the underlying scale using traditional methods is available from previous studies. Using these data an analyst poses the following questions:

1. How many genetic markers influencing the trait, if any, are detected, and how is this detection accomplished?
2. What proportion of the variance on the underlying scale is captured by the regression on markers?
3. How do genotype profiles at the detected loci differ between two individuals with low and high estimated probabilities of showing the disease?
4. Should the predictor include the "significant" markers only or should all the markers be incorporated? What justifies the decision?
5. If incorporation of genetic marker information is useful for prediction, how should the model be tested/validated?
6. How can a measure of uncertainty be attached to the prediction?

Data consist of $5,000$ individuals genotyped for $15,000$ genetic markers with observed phenotypic binary records $y_i = (0, 1)$. The $5,000$ individuals are randomly divided into training (y_t) and validating data (y_v), $2,500$ in each. Training data y_t are analysed with the following hierarchical spike and slab model:

$$\begin{aligned} \Pr\left(y_{t,i} = 1 | b, x_{t,i}\right) &= \Phi\left(\mu + x'_{t,i} b\right), \\ y_{t,i} | b, x_{t,i} &\sim Bin\left(1, \Phi\left(\mu + x'_{t,i} b\right)\right), \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, μ is an unobserved mean, $x'_{t,i}$ is the ith row of an observed matrix x of genetic markers of dimension $2,500 \times 15,000$ and b is the unobserved vector of $15,000$ marker effects. Marker effects are assumed independently drawn from the mixture prior

$$b_j|\sigma_b^2, \pi \sim N\left(0, \sigma_b^2\right)\pi + \Delta_0\left(1-\pi\right),$$

where the scalar σ_b^2 is the uncertainty variance of a marker effect drawn form the normal distribution, π is the a priori proportion of markers originating from the normal component in the population, and Δ_0 is a point probability mass centered at zero. The final layers of the hierarchy are

$$\begin{aligned} \pi|\alpha, \beta &\sim Be\left(\alpha, \beta\right), \\ \sigma_b^2 &\sim \text{scale inverted } \chi^2\left(S_b, v_b\right), \end{aligned}$$

where $Be\left(\alpha, \beta\right)$ represents a beta distribution with user-tuned parameters α and β chosen to generate a prior mode in the proximity of 0.001, S_b is the scale of the scale inverted chi-square prior distribution of σ_b^2 and v_b the degrees of freedom. The latter is set equal to 4.1 and the scale is set equal to 0.721 to generate a prior modal value of σ_b^2 approximately equal to 0.14.

The spike and slab model is implemented using the McMC algorithm described on page 380.

The data generating model or true model, unknown to the analyst, is as follows. At the level of the liability u_i, for individual i,

$$u_i = \mu + x'_{Qi}b_Q + a_i + \varepsilon_i, \quad i = 1, 2, \ldots, n = 5,000, \tag{1}$$

where the mean μ is chosen so that $\Pr\left(Y_i = 1|\mu\right) = 0.05$, x'_{Qi} is the ith row of a matrix x_Q of QTL (causal) genotypes of dimension $n \times (nqtl = 10)$, b_Q is the $nqtl \times 1$ vector of additive genetic effects of the QTL genotypes, a_i is the effect of additive genetic contributions from an unknown number of causal loci, and $\varepsilon_i \sim N\left(0, 1\right)$ is the effect of a normally distributed random residual term. When individuals are nominally unrelated, a_i and ε_i can be grouped in a single term $e_i \sim N\left(0, \left(\left(1-r_Q\right)\sigma_a^2 + 1\right)\right.$, where $\sigma_a^2 = 0.5$ is the additive genetic variance at the level of the liability and r_Q is the true proportion of the total additive genetic variance captured by the $nqtl$ QTL causal loci. This additive genetic variance has a contribution r_Q from the $nqtl$ QTL, and a contribution $(1-r_Q)$ from the unobserved additive genetic component a. The last three terms of the right hand side of (1) are assumed to be independent.

The size of the $nqtl$ effects of the vector b_Q is chosen to generate an additive genetic variance equal to $r_Q\sigma_a^2$.

The elements of the matrix x_Q of QTL genotypes are part of the matrix x of genetic markers. The $nqtl$ loci are randomly chosen among the $15,000$ markers and assigned as QTL loci. The elements of the $n \times nqtl$ QTL loci are independently drawn from a binomial distribution $Bin\left(2, p = 0.05\right)$. Those from the $n \times (15,000 - nqtl)$ non-QTL

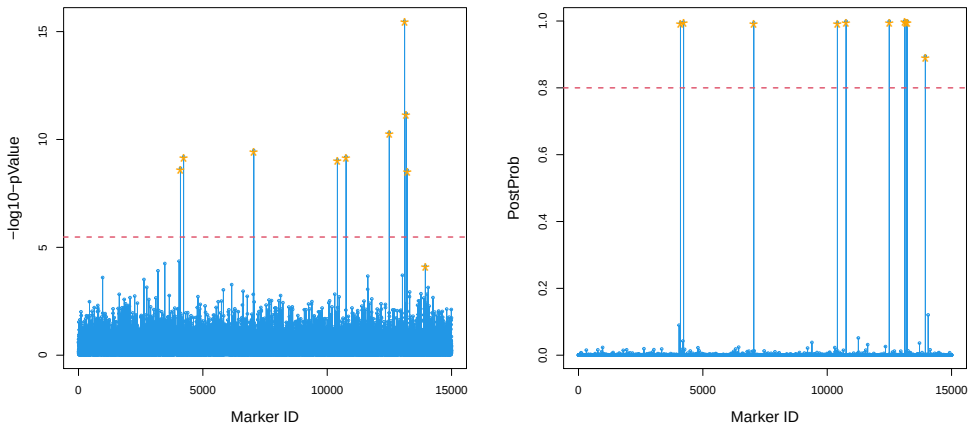

Figure 1: Detection of promising marker genotypes; 2 500 training observations. Left: GWAS with a Bonferroni correction; 9 out of 10 markers detected; 0 true false discoveries. The horizontal red line is the Bonferroni threshold corresponding to $-log_{10}(0.05/15,000)$. Right: spike and slab model; 10 out of 10 markers detected; 0 true false discoveries. The horizontal red line is the arbitrary threshold corresponding to a posterior probability equal to 0.8. Orange stars represent the 10 true genes.

marker loci, are independently drawn from a binomial distribution $Bin\,(2, p = 0.5)$. Thus, genetic markers are present at intermediate frequencies and QTL loci at low ($p = 0.05$) frequencies.

The example is in two parts that differ in the true data generating model. In Part I $r_Q = 1$ and the $nqtl = 10$ QTL loci explain 100% of the additive genetic variance σ_a^2. In this case the 10 causal marker loci can be detected despite their low frequency in the sample of $2,500$ training observations. This illustrates a case where the experiment has high power, the prediction ability of the model is high and results are relatively clear and uncontroversial. In Part II $r_Q = 0.4$ and the $nqtl = 10$ QTL loci explain 40% of the additive genetic variance σ_a^2. This leads to a lower probability of detection of the QTL loci and the prediction ability of the model is compromised.

Before addressing the six questions posed above, a little intuition for the problem at hand can be obtained as follows. Consider one of the 10 causal marker loci. Assuming Hardy-Weinberg equilibrium the three possible genotypes $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ occur with expected frequencies $(0.95)^2$, $2\,(0.95)\,(0.05)$ and $(0.05)^2$. The expected number of the three genotypes in a sample of $2,500$ individuals is $2\,256.25$, 237.5, 6.25. The three conditional probabilities of the event occurring, given the population mean $\mu = -1.64$ and the QTL locus are $\Phi\,(\mu + bx_0)$, $\Phi\,(\mu + bx_1)$, $\Phi\,(\mu + bx_2)$. In Part I of the exercise, where the 10 QTL loci explain 100% of the additive genetic variance of the trait on the underlying scale, $b = 0.72$. This translates into the following three values for the three conditional probabilities: 0.05, 0.17, 0.42, with a posterior standard deviation of the largest

of these three probabilities, accounting for uncertainty of b, of the order of 0.06. Most of the information in the data for drawing inferences about these probabilities is contributed by individuals with genotypes x_0 and x_1 (because only 6 individuals, approximately, are expected to carry genotype x_2). Very few cases are expected with individuals carrying genotypes $x = x_2$ at more than one or two of their ten QTL loci. In Part II of the exercise where the 10 QTL loci explain 40% of the additive genetic variance of the trait on the underlying scale, the allele substitution effect $b = 0.46$. The three conditional probabilities of the event occurring are now 0.05, 0.12 , 0.23. The posterior standard deviation of the largest of these three probabilities, accounting for uncertainty of b, is of the order of 0.10. This has a strong impact on QTL detection and on the evaluation of the model's predictive ability.

Part I

1. The results of a GWAS with a Bonferroni adjustment fitting one marker at a time for all 15 000 genetic markers are shown in the left panel of Figure 1. The right panel displays similar information based on the spike and slab model. GWAS detects 9 out of the 10 QTL with zero true false positives (a detail unknown to the analyst); the spike and slab model detects all the 10 QTL with zero true false positives. This information is supplemented with the computation of false discovery rates using the Benjamini and Hochberg (FDR-BH) approach and the fully Bayesian approach. Implementation of the FDR-BH rule, setting q equal to 0.10 (see page 338) leads to a discovery set equal to 12 markers, with 2 true (unknown to the analyst) false positives, and 10 true discoveries. The fully Bayesian approach, using the right panel of Figure 1 to guide the choice of the discovery set based on the estimate of the posterior probability that the marker is drawn from the slab, equal to 0.8, leads to a discovery set of 10 markers, with 0 true false positives. The estimate of the posterior mean of the Bayesian FDR is 0.01, with a 95% posterior interval equal to $(0.00; 0.1)$. The Bayesian result indicates that the estimated average of the posterior distribution of the number of false discoveries in the set of 10 markers that constitute the discovery set is $0.01 \times 10 = 0.1$. With the information available the analyst decides to build a prediction model incorporating the discovery set based on the 10 genetic markers detected by the spike and slab model.

2. The estimated posterior mean and 95% posterior interval of the genomic variance on the underlying scale based on the spike and slab model using all markers is 0.56, $(0.43, 0.72)$. This is in quite good agreement with the true value equal to 0.5. The posterior distribution of the proportion of variance on the underlying scale explained by all the 15 000 markers (or genomic heritability) has mean and 95% posterior interval equal to $0.56/1.56 = 0.36$ $(0.30; 0.42)$. The estimated posterior mean and 95% posterior interval of the genomic variance on the underlying scale based on the 10 detected markers is 0.52, $(0.41, 0.64)$, and the corresponding figures for the genomic heritability are 0.34 $(0.29; 0.39)$.

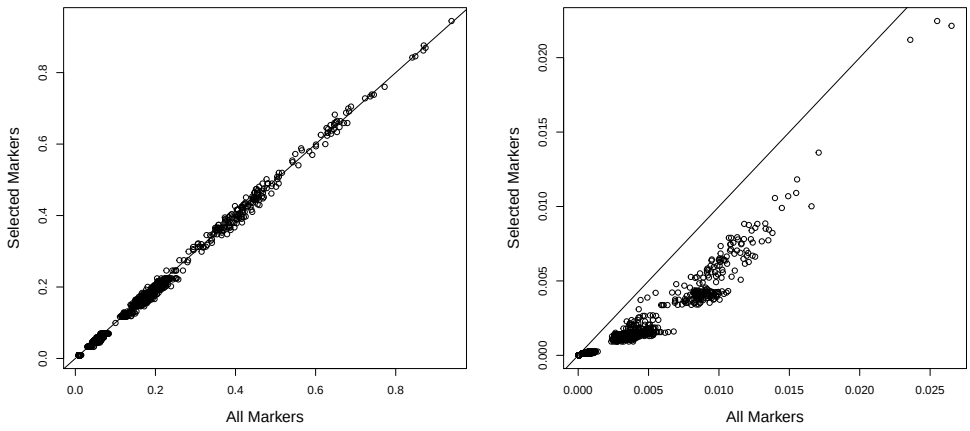

Figure 2: Left: McMC estimates of posterior means of $\Pr(Y_i^* = 1|x_i, y_t)$ based on a model that includes either 10 detected markers or the complete marker panel. Right: McMC estimates of posterior variances of $\Pr(Y_i^* = 1|x_i, y_t)$ based on a model that includes either 10 detected markers or the complete marker panel. True model: 10 QTL account for 100% of the genomic variance.

3. The spike and slab model outputs Monte Carlo estimates of $\Pr(Y_i^* = 1|x_i, y_t)$; these are posterior predictive predictions, accounting for uncertainty of b. The posterior mean (95% posterior interval) of the highest scoring probability including all markers (that happens to correspond to individual 889) is 0.94 $(0.78; 0.99)$. The posterior mean (95% posterior interval) of the lowest scoring probability including all markers (that happens to correspond to individual 2300) is 0.008 $(0.001; 0.018)$. The genotype profiles at the detected marker loci for 889 is (before column centring)

   $$1\,0\,0\,2\,0\,0\,1\,0\,1\,1$$

   and for 2300

   $$0\,0\,0\,0\,0\,0\,0\,0\,0\,0.$$

   Individual 889 has one copy of a "disease causing gene" at 4 of the 10 detected loci, and two copies at one locus. Individual 2300 does not have any "disease causing alleles". An individual that has an estimated probability $\Pr(Y_i^* = 1|x_i, y_t) = 0.085$, corresponding to the mean of the validating data, has a genotypic profile at the detected marker loci

   $$0\,0\,0\,0\,0\,0\,1\,0\,0\,0.$$

   This result could have implications for personalised medicine.

4. Figure 2, left panel, shows good agreement between predictions based on all markers and those based on detected markers only. This is hardly surprising in this example

where the QTL account for all the genomic variance on the underlying scale of the trait. However the right panel indicates how the variance of the predictor increases with the incorporation of a large number of covariates. Each point represents either the posterior mean (left panel) or the posterior variance (right panel) of the prediction for a particular individual, for all the 2 500 individuals in the validating data. In the case of this example, the conclusion is that only the detected markers should be included in the construction of the predictor. This may not be the case when adding more markers contribute with relatively more information than noise.

5. To study the model's prediction ability several exploratory measures are investigated. Most of these originate from various outputs of the McMC implementation of the spike and slab model. The first two are the validating mean squared error (proportion of misclassifications) using the Bayes classifier to transform probabilities into values 0 or 1, generating in this way predicted data $\hat{Y}$, and the Brier score. Both are compared to the validating mean squared error and Brier score based on the null model $\Pr(Y = 1|\mu)$ as benchmark, where μ is replaced by the estimate of the posterior mean of μ. These two criteria are implemented using predicted probabilities computed with all the 15 000 markers, or with the 10 selected markers only. In addition, the same criteria are applied using the highest and lowest predicted probabilities generated with the selected markers. Finally, logscores evaluated at the estimated mean of the posterior distributions of the predicted probabilities are calculated based on either all the 15 000 markers, or with the 10 selected markers only.

	Misclassification	Brier score
All markers	0.078	0.062
Detected markers	0.078	0.062
Highest prob	0	0.017
Lowest prob	0	0.005
Null model All	0.085	0.078

Table 1: Point estimates of validating misclassification rates and of validating Brier scores incorporating (i) all 15 000 markers; (ii) only the 10 detected markers; (iii) using a subset of the highest or (iv) lowest predicted probabilities $\Pr(\hat{Y}_v = 1|x, y_t)$ computed using the 10 detected markers, and (v), using the null model with all the validating data points $\Pr(\hat{Y}_v = 1|\hat{\mu}, y_t)$, where $\hat{\mu}$ is the estimate of the posterior mean of μ.

Results are displayed in Table 1. The same qualitative conclusions are drawn from the misclassification rates or the Brier score. The point estimates of validating mean squared error are computed using

$$MSE_v = \frac{1}{n_v} \sum_{i=1}^{n_v} (y_{v,i} - \hat{y}_{v,i})^2 . \tag{2}$$

In this expression $n_v = 2\,500$, $y_{v,i}$ is the ith validating datum, $\hat{y}_{v,i}$ is its prediction that takes values 0 or 1 according to $I\left[\hat{\Pr}\left(y_{v,i} = 1|x_i, y_t\right) > 0.5\right]$, where $I\left(\cdot\right)$ is the indicator function that takes the value 1 if the argument is satisfied and 0 otherwise, y_t is the training data and $\hat{\Pr}\left(y_{v,i} = 1|x_i, y_t\right)$ is the McMC estimate of the posterior mean $E_{b|x_i,y_t}\left[\Pr\left(y_{v,i} = 1|b, x_i, y_t\right)\right] = \Pr\left(y_{v,i} = 1|x_i, y_t\right)$ for datum i. The Brier score replaces $\hat{y}_{v,i}$ in (2) with the McMC estimates of the posterior means $\hat{\Pr}\left(y_{v,i} = 1|x_i, y_t\right)$. Predictions based on the complete marker panel or on the detected markers only (first two rows of the table) lead to the same estimates of validating mean squared error and Brier score. On the face of it, parsimony would dictate choosing a predictor based on the 10 detected markers. However, the performance of the model using the complete set of 2 500 predictions is only marginally superior to the performance of the null model (MSE_v equal to 0.078 and 0.085, respectively). This is typically the case with binary data where the event occurs with very low probability. More convincing evidence in favour of the model's predictive ability can be sought by studying the misclassification rates and the Brier score based on the model that incorporates the 10 detected markers, choosing extreme values of the estimated predictions. The misclassification rates and Brier score among the 10 highest and lowest predictions are 0 and close to the minimum possible values, respectively (rows 3 and 4 in the table). The 10 validating data points corresponding to these 10 highest and lowest predictions are all equal to 1 and to 0, respectively.

The average of the ten highest estimated predictions $\hat{\Pr}\left(y_v = 1|x, y_t\right)$ is 0.87 and the average of the ten lowest is 0.0091. These are respectively 10.2 times larger and 9.3 times lower than the prediction based on the null model.

The logscore is computed using

$$\sum_{i=1}^{n_v} y_{v,i} \log\left(\hat{P}_i\right) + \left(1 - y_{v,i}\right) \log\left(1 - \hat{P}_i\right), \tag{3}$$

where $\hat{P}_i = \hat{\Pr}\left(y_{v,i} = 1|x_i, y_t\right)$, the average estimated McMC estimate of the posterior predictive probability for the ith validating datum. When the validating datum $y_{v,i} = 1$, if $\hat{P}_i$ is close to 1, the record makes a minimal contribution to the logscore. The same occurs when the validating datum $y_{v,i} = 0$ and $\hat{P}_i$ is close to 0. Discrepancies between the observed validating data and their predictions lead to relatively large negative contributions. In the present scenario, the logscores in a model that uses all the 15 000 markers, the 10 detected markers and the null model, are respectively -536.7, -535.9 and -760.2. These figures agree with those in Table 1, supporting the claim that genetic markers contribute to the model's prediction ability, and that there is no extra benefit to include other than the 10 detected markers in the construction of the predictor. An estimate of the complete posterior predictive distribution of the logscores can be obtained using the draws $P_i^{[t]}$ at round t of the McMC algorithm.

As a further analysis of the predictive ability of the model constructed with the 10 detected loci, using output from the McMC algorithm, one can obtain an estimate

of the conditional probability

$$\Pr\left[\left(\Pr\left(\hat{y}_v = 1|x, y_t\right) > 2\Pr\left(\hat{y}_v = 1\right)\right)|y_v = 1\right]. \tag{4}$$

In words, among those individuals in the validating data that show the disease ($y_v = 1$), what proportion of the estimated posterior predicted probabilities ($\Pr\left(\hat{y}_v = 1|x, y_t\right)$) is larger than twice the probability of observing disease in an individual randomly sampled from the population ($2\Pr\left(\hat{y}_v = 1\right)$)? The answer is 0.72. Similarly, given that the true state of nature is $y_v = 0$, the model predicts a disease outcome with probability

$$\Pr\left[\left(\Pr\left(\hat{y}_v = 1|x, y_t\right) > 2\Pr\left(\hat{y}_v = 1\right)\right)|y_v = 0\right],$$

equal to 0.18, smaller then (4) by a factor of 4.

With the evidence presented one can recommend using the model as a prediction tool on an exploratory basis. Sharper inferences should be possible as further data are collected.

6. The McMC algorithm that implements the spike and slab model outputs draws from the posterior distribution $\Pr\left(y_{v,i} = 1|x_i, \mu, b, y_t\right) = \Phi\left(\mu + x_i'b\right)$, where (μ, b) is a draw from $[\mu, b|y_t]$. The variance among these draws is a consistent estimator of the posterior variance of $\Phi\left(\mu + x_i'b\right)$, that reflects the posterior uncertainty of (μ, b) propagating on to the posterior distribution of $\Phi\left(\mu + x_i'b\right)$. As an illustration, consider again the posterior mean (95% posterior interval) of the highest scoring probability including only the 10 detected markers (that happens to correspond to individual 889, the same as in point 3 above). With only 10 markers included in the construction of $\Phi\left(\mu + x_i'b\right)$, this probability is 0.94 (0.83; 0.99); the posterior interval is a little narrower than the value (0.78; 0.99) obtained using all markers.

Part II

1. The results of the GWAS and of the spike and slab model are shown in Figure 3, on the left and right panels, respectively. Two out of the 10 markers reach beyond the Bonferroni threshold and both are true positives (left subfigure). The four highest posterior probabilities that markers have an effect on the trait estimated with the spike and slab model (right subfigure, all true positives) are 0.51, 0.15, 0.35 and 0.48. Implementation of the Benjamini and Hochberg false discovery rate algorithm (FDR-BH), setting the expected proportion of false positives q equal to 0.15, leads to a discovery set of 4 markers (all happen to be true discoveries). The result is unchanged using $q = 0.5$. These are the same 4 markers that reach the highest probabilities in the analysis based on the spike and slab model. Implementation of the fully Bayesian FDR with a discovery set defined by those 4 markers that reach estimated probabilities larger than 0.15, outputs an estimate of the false discovery rate with a mean of 0.6 and an estimated 95% posterior interval equal to (0.00; 1.00).

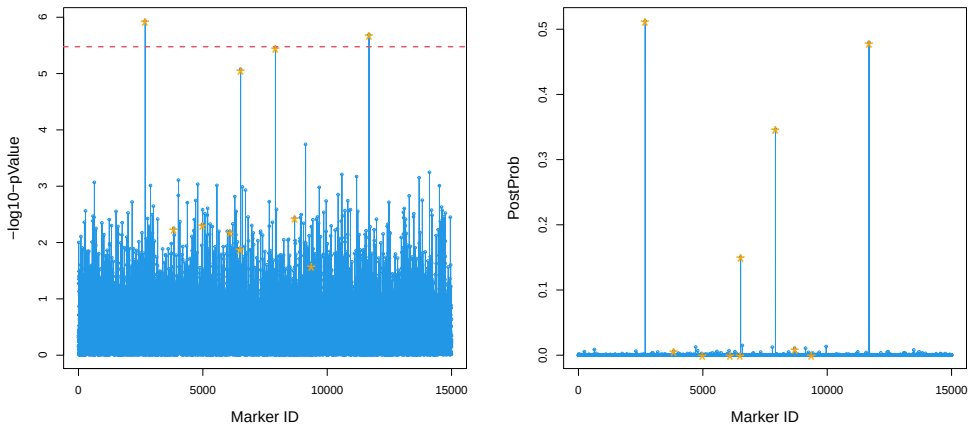

Figure 3: Detection of promising marker genotypes; 2 500 training observations. Left: GWAS with a Bonferroni correction; 2 out of 10 markers detected (marker 7 919 on the border of the detection threshold); 0 true false discoveries. The horizontal red line is the Bonferroni threshold corresponding to $-log_{10}(0.05/15,000)$. Right: spike and slab model; Orange stars represent the 10 true genes.

The large value of the estimated FDR is due to the low estimates of the posterior probability. For example, if the discovery set only includes the marker with highest estimate of the posterior probability (this is marker 2 681, with estimate equal to 0.51), the estimate that a false discovery has been made is $1-0.51=0.49$. The large estimated uncertainty is due to the poor information content in the data to estimate the marker effects. This leads to strong fluctuations in the estimates of the posterior probabilities, that propagate on to the estimates of FDR. With these shortcomings in mind, the analyst decides to construct a model based on the 4 detected markers.

2. The estimated posterior mean and 95% posterior interval of the genomic variance on the underlying scale based on the spike and slab model using all the 2 500 markers is 0.04 $(0.00; 0.10)$. The corresponding figure based on the 4 detected markers is 0.03 $(0.00, 0.08)$. The analyst does not feel encouraged by these figures bearing in mind the prior estimate of genomic variance equal to 0.5. The true (unknown to the analyst) proportion of the additive genetic variance explained by the 10 QTL is $0.5 \times 0.4 = 0.2$. Note that assuming independent loci, 4 QTL should generate a genomic variance equal to 0.08 square units; this is the value (unknown to the analyst) against which the estimate 0.03 should be compared.

3. The extreme values (and their 95% posterior intervals) of the estimated $E_{b|y_t}(\Pr(Y_i^*=1|b,x,y_t)=\Pr(Y_i^*=1|x,y_t)$ using the 4 selected marker genotypes are 0.24 $(0.07; 0.54)$ and 0.069 $(0.05; 0.09)$, where the confidence interval reflects posterior uncertainty of

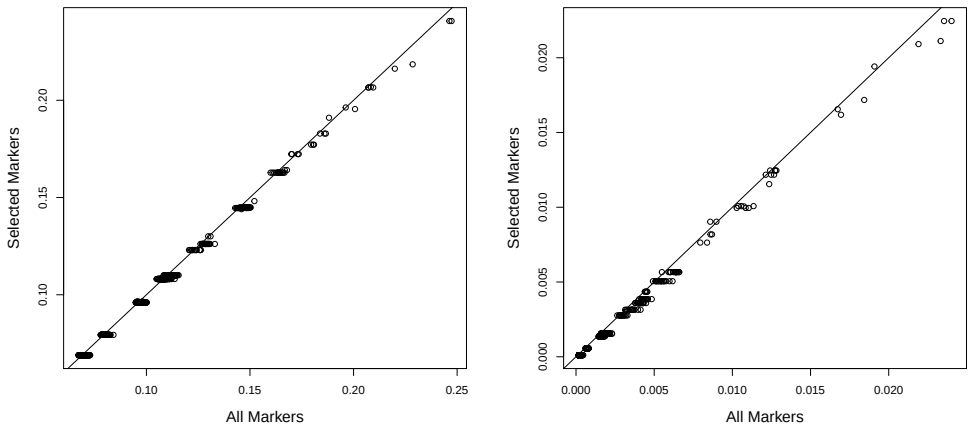

Figure 4: Left: McMC estimates of posterior means of $\Pr(Y_i^* = 1|x_i, y_t)$ based on a model that includes either 4 detected markers or the complete marker panel. Right: McMC estimates of posterior variances of $\Pr(Y_i^* = 1|x_i, y_t)$ based on a model that includes either 4 detected markers or the complete marker panel. True model: 10 QTL account for 40% of the genomic variance.

$(b|y_y)$. The genotypes at the 4 detected loci for these extreme values are

$$2\ 0\ 0\ 1$$

and

$$0\ 0\ 0\ 0,$$

respectively. On the other hand, the genotype combination of the 4 detected loci in the data that generates the highest prediction $\Pr(Y_i^* = 1|x, y_t)$ is

$$1\ 2\ 0\ 1.$$

However the estimated value (95% posterior interval) $\hat{\Pr}(Y_i^* = 1|x = (1, 2, 0, 1), y_t)$ is 0.21 $(0.07; 0.62)$, smaller than the estimated prediction equal to 0.24 based on the genotype with lower gene content $x = (2, 0, 0, 1)$. In the case of genotype $x = (1\ 2\ 0\ 1)$, a histogram (not shown) of the draws from the element of $(b|y_t)$ corresponding to the second detected locus (with gene content 2) reveals that most of the probability mass is concentrated near $b = 0$. This happens to be one of the 4 detected QTL with smallest probability of affecting the trait (see right panel of Figure 3). Therefore this detected locus combination, despite its high overall gene content equal to 4, contributes little to the estimated $\hat{\Pr}(Y_i^* = 1|b, x, y_t)$. When data do not contain enough information to infer accurately marker substitution effects b, the prediction ability of a model built on a panel of selected marker genotypes is compromised.

4. Figure 4, left panel, shows again good agreement between predictions based on all markers and those based on the 4 detected markers only. The right panel shows that the variance of the predictor increases with the incorporation of the complete set of covariates, although the effect is less marked than in Part I. For example, the variance of the predictor corresponding to the highest estimated prediction based on the 4 detected markers is

$$Var_{b|y_t}\left[\Pr\left(Y_i^* = 1|b, x_i = (2\ 0\ 0\ 1), y_t\right)\right] = 0.022, \quad i = 255,$$

whereas the value based on the 15 000 markers is 0.024. As in Part I, the conclusion must be that only the 4 detected markers should be used in the construction of the predictor.

5. A striking difference between the behaviour of the prediction models in Part I and Part II is the range of the predicted probabilities. Predictions in Part I range from close to 0 to over 0.9; in Part II, from close to 0 to a little over 0.2 (see left panels of Figures 2 and 4). These observations together with the proportion of $1's$ in the sample of validating data equal to 0.089 have a strong influence on the results that follow.

	Misclassification	Brier score
All markers	0.089	0.080
Detected markers	0.089	0.080
Highest prob	0.500	0.337
Lowest prob	0.100	0.091
Null model All	0.089	0.081

Table 2: Point estimates of validating misclassification rates and of validating Brier scores incorporating (i) all 15 000 markers; (ii) only the 4 detected markers; (iii) using a subset of the highest predicted probabilities $\Pr(\hat{Y}_v = 1|x, y_t)$ computed using the 4 detected markers; (iv) using a subset of the lowest predicted probabilities; (v) using the null model including all the validating data points $\Pr(\hat{Y}_v = 1|\hat{\mu}, y_t)$, where $\hat{\mu}$ is the estimate of the posterior mean of μ.

The first two rows of Table 2 indicate that predictions based on all the 15 000 markers or on the 4 detected markers lead to the same estimates of validating mean squared error and Brier score. Predictions based on the null model using all validating records, (last row of the table) that in the present case amounts to setting all the 2 500 predictions equal to zero, lead to the same estimates of validating mean squared error and Brier score. These initial results suggest that incorporating information on marker genotypes, (whether it is 4 markers or 15 000) does not contribute to the model's predictive ability. One can seek more discriminatory power by looking at the same measures of prediction ability among the extreme highest and lowest predictions, based on the 4 detected markers (rows 3 and 4 in the table). The highest

ten observed validating data points consist of 5 $1's$ and 5 $0's$. The corresponding predictions using the Bayes classifier (that minimises the overall proportion of misclassifications) are all equal to 0. This results in the misclassification in the table equal to 0.5. The average of the 10 highest predicted probabilities is approximately 0.21. The Brier score using this average is $0.5(1-0.21)^2+0.5(0-0.21)^2 = 0.334$, and using the individual predicted values yields the figure 0.337 in the table. A similar exercise holds for the figures corresponding to the lower extreme, in the 4th row of the table. The lowest ten observed validating data points consist of a single 1 and 9 $0s$ and the predictions based on the Bayes classifier are all equal to 0. The 10 lowest predicted probabilities are all approximately 0.069. This leads to a misclassification rate among the lowest predictions equal to 0.1 and to a Brier score equal to $0.1(1-0.069)^2 + 0.9(0-0.069)^2 = 0.091$ (row 4 in the table).

The misclassification rates and the Brier score indicate that the model's ability to predict single observations is poor. However, the model performs better to select a subset of the data with a proportion of affected individuals that deviate markedly from the mean proportion in the population. In the case described above, the highest 10 predicted probabilities out of 2 500 correspond to a proportion of affected individuals in the validating data equal to 50%; the corresponding proportion among the lowest predicted probabilities is 10%, a little larger than the mean proportion of affected cases in the validating data equal to 8.9%. These figures are respectively 24% and 6% for the extreme 100 selected probabilities out of 2 500.

The logscores in a model that uses all the 15 000 markers, the 4 detected markers and the null model, are respectively -728.7, -729.3 and -754.00. The result suggests that predictions based on the 4 detected markers perform as well as those based on all 15 000 markers, and that both do a little better than the null model.

The final test of the prediction ability of the model incorporating the 4 detected markers is based on calculating what proportion of the estimated posterior predicted probabilities ($\Pr(\hat{y}_v = 1|x, y_t)$) is larger than twice the probability of observing disease in an individual randomly sampled from the population ($2\Pr(\hat{y}_v = 1)$)? The answer now is 0.036. Similarly, given that the true state of nature is $y_v = 0$, the model predicts a disease outcome with probability

$$\Pr\left[\left(\Pr\left(\hat{y}_v = 1|x, y_t\right) > 2\Pr\left(\hat{y}_v = 1\right)\right)|y_v = 0\right],$$

equal to 0.004, smaller than 0.036 by a factor of 9. The model does much better at assigning low probabilities of disease to healthy individuals, than to assign high probabilities of disease to unhealthy individuals.

With the evidence presented one can only be hesitant to recommend using the model as a prediction tool for early diagnosis. An alternative design based on case-control studies could mitigate some of the shortcomings arising from the low frequency of the disease among randomly sampled data and from the small proportion of genomic variance accounted for by the marker information. Adoption of a case-control design may require accounting for non-random sampling.

6. Estimates of the 2 500 complete marginal posterior distributions of the probabilities for each validating datum can be easily obtained from the output of the McMC algorithm. For example, for the largest and smallest probabilities, using either the complete marker panel or the detected markers, the estimated means and variances (in brackets) are as follows. For the largest and smallest, all markers: 0.247 (0.0242) and 0.067 (1.94×10^{-4}). For the largest and smallest, detected markers: 0.241 (0.0224) and 0.069 (1.05×10^{-4}).

References