Methods S1

**Whole Genome Regression & Prediction of Censored Traits Using and Extended Version of the Bayesian LASSO**

In our WGP model, we describe the conditional expectation function of the response (, *i=1,…,n*) as the sum of individual-specific means () and of a model residual () which is assumed to be a normal random variable with mean zero and variance ; therefore . Here,

, (1)

is a linear predictor that has three components: , an effect common to all subjects; , a regression component accounting for the effects of non genetic covariates (sex, smoking and BMI covariates in our application) ; and , a regression on SNP genotypes  where  counts the number of copies of the least frequent allele at the *jth* SNP. Combining the above-described assumptions we obtain the likelihood function un-censored data points

 . (2)

and that of surviving subjects

 , (3)

where  represent age at censoring of the *ith* individual and  is the cumulative distribution function of a standard normal random variable. Expression (3) represents the likelihood function of a right-censored observation. Similar expression can be derived for left and interval censored observations; herein we focus on right-censored observation only.

Combining equations (1)-(3) and under conditional independence, we obtain the conditional density of the data given effects and the residual variance:

|  |  |
| --- | --- |
| ***Likelihood:*** |  (4) |

where ,  ,  and  are vectors of data effects and age at censoring, respectively, and and  indicate that the product runs over the set of un-censored and censored records, respectively.

Bayesian inference is based on the posterior density of the unknowns  given the data. Following Bayes’ theorem, this is proportional to the conditional density of the data, given by (4), times the prior density of model unknowns. Here, we structure the prior density using a modified version of the Bayesian LASSO [1]. We extend this model to accommodate censoring, according to (4), as well as effects other than those of markers. Specifically, we assign independent vague prior densities to the intercept () and to the effects of sex, smoking and BMI (). This treatment yields estimates of the effects of these non-genetic factors that are comparable to those obtained with likelihood-based methods. Following standard assumptions of Bayesian regression models, the residual variance is assigned a scaled-inverse Chi-square density, , with degree of freedom and scale parameter  and , respectively. The prior density of marker effects is structured using the specification used in the Bayesian LASSO of Park and Casella [1], here,

,

were:  is a normal density assigned to the effect of the *lth* marker, centered at zero and with marker-specific prior variance . In the next level of the hierarchy, the prior variance of marker effects are assigned IID exponential priors, , each indexed by a regularization parameter, , which controls the shape of this prior and, with this, the amount of shrinkage of estimates of marker effects. In the last level of the hierarchy the regularization parameter is assigned a Gamma prior density. Such treatment allows inferring this important parameter from the data. Collecting the assumptions above described, the joint prior density becomes:

***Prior***:  (5)

Expression (5) is indexed by a set of four hyper-parameters, a discussion of how these can be chosen is given in Perez et al.[2].

Following Bayes’ theorem the posterior density is proportional to the product of the conditional density of the data (4) and the prior (5); therefore,



This posterior density does not have closed form; however, using data-augmentation we can draw samples from the posterior density using a Gibbs sampler. A modified version of the BLR package [27] that handles right, left and interval censored was developed and it is available upon request to the corresponding author.