University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Wang, T;Chen, YPP;Bowman, PJ;Goddard, ME;Hayes, BJ

Title:
A hybrid expectation maximisation and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping

Date:
2016-09-21

Citation:
Wang, T., Chen, Y. P. P., Bowman, P. J., Goddard, M. E. & Hayes, B. J. (2016). A hybrid expectation maximisation and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping. BMC Genomics, 17 (1), https://doi.org/10.1186/s12864-016-3082-7.

Persistent Link:
https://hdl.handle.net/11343/256666

License:
CC BY

Wang et al. BMC Genomics (2016) 17:744
DOI 10.1186/s12864-016-3082-7

BMC Genomics

RESEARCH ARTICLE Open Access

A hybrid expectation maximisation and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping

Tingting Wang[1,2,3*], Yi-Ping Phoebe Chen[1], Phil J. Bowman[2,3,4], Michael E. Goddard[2,3,5] and Ben J. Hayes[2,3,6]

Abstract

Background: Bayesian mixture models in which the effects of SNP are assumed to come from normal distributions with different variances are attractive for simultaneous genomic prediction and QTL mapping. These models are usually implemented with Monte Carlo Markov Chain (MCMC) sampling, which requires long compute times with large genomic data sets. Here, we present an efficient approach (termed HyB_BR), which is a hybrid of an Expectation-Maximisation algorithm, followed by a limited number of MCMC without the requirement for burn-in.

Results: To test prediction accuracy from HyB_BR, dairy cattle and human disease trait data were used. In the dairy cattle data, there were four quantitative traits (milk volume, protein kg, fat% in milk and fertility) measured in 16,214 cattle from two breeds genotyped for 632,002 SNPs. Validation of genomic predictions was in a subset of cattle either from the reference set or in animals from a third breeds that were not in the reference set. In all cases, HyB_BR gave almost identical accuracies to Bayesian mixture models implemented with full MCMC, however computational time was reduced by up to 1/17 of that required by full MCMC. The SNPs with high posterior probability of a non-zero effect were also very similar between full MCMC and HyB_BR, with several known genes affecting milk production in this category, as well as some novel genes. HyB_BR was also applied to seven human diseases with 4890 individuals genotyped for around 300 K SNPs in a case/control design, from the Welcome Trust Case Control Consortium (WTCCC). In this data set, the results demonstrated again that HyB_BR performed as well as Bayesian mixture models with full MCMC for genomic predictions and genetic architecture inference while reducing the computational time from 45 h with full MCMC to 3 h with HyB_BR.

Conclusions: The results for quantitative traits in cattle and disease in humans demonstrate that HyB_BR can perform equally well as Bayesian mixture models implemented with full MCMC in terms of prediction accuracy, but with up to 17 times faster than the full MCMC implementations. The HyB_BR algorithm makes simultaneous genomic prediction, QTL mapping and inference of genetic architecture feasible in large genomic data sets.

Background

Genomic prediction of genetic merit, using SNP markers, is now routinely used in animal and plant breeding to identify superior breeding individuals and so accelerate genetic gain [1–3]. Genomic prediction methodology is also increasingly used in human disease studies for the inference of genetic architecture, the identification of causal mutations (QTL mapping), and prediction of disease risk [4–8].

Genomic predictions are often implemented using linear prediction models, especially best linear unbiased prediction (BLUP) or Genomic BLUP (GBLUP), which assume that all the SNPs contribute small effects to the trait and these effects are derived from a normal distribution [1, 4, 9]. While BLUP based genomic predictions have certainly been successful in increasing genetic gain

* Correspondence: t22wang@students.latrobe.edu.au
[1]School of Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC, Australia
[2]Biosciences Research, Department of Economic Development, Jobs, Transport and Resources, Bundoora, Melbourne, VIC, Australia
Full list of author information is available at the end of the article

© 2016 The Author(s). **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

in livestock and crops [10, 11], this approach does have some limitations. One limitation is that the prediction accuracy does not persist well across multiple generations, because the severe shrinkage in these models results in the effect of causative mutation being "smeared" across many markers encompassing long chromosome segments – in other words a linear combination of effects of a large number of markers is used to capture the effect of a QTL. After several generations, the association between markers and QTL might be broken down by recombination, thereby reducing prediction accuracy. The smearing of effect of causative mutations across many SNP also results in imprecise QTL mapping with BLUP methods.

To address these problems, Bayesian mixture models (nonlinear prediction e.g. Bayes A, B, C, and R) [1, 6, 12–15] assume non-normal prior distributions of SNP effects. One example of a flexible approach, BayesR [14] defines a mixture model with SNP effects following a mixture of four normal distributions with zero, very small, small and moderate variances. In practice, the prediction accuracy of Bayesian mixture models (including BayesR) has been shown to be equal or superior to that of GBLUP for both human diseases and dairy cattle milk production traits [15–25].

In addition to the prediction of breeding values and future phenotypes using genotype data, Bayesian models (such as BayesR) can be used to map the causal polymorphisms (quantitative trait loci or QTL). For this purpose they have some advantages over traditional single SNP regression, which is commonly used to analyse genome wide association studies (GWAS) [16–24]. Single SNP regression fits one SNP at a time as a fixed effect and tests the significance of the association between the SNP and the trait, while ignoring all other SNPs. To protect against performing multiple tests, stringent P-values ($P < 5^{*}10^{-8}$) are used. This method of analysis has three limitations:1) The effect of those SNPs declared significant is usually over-estimated; 2) multiple SNPs in linkage disequilibrium with the same QTL may show an association with the trait leading to imprecision in mapping the QTL; 3) many QTL are not detected at all because no SNP reaches the stringent *P*-value for association with the trait. By comparison, Bayesian mixture models fit all SNPs simultaneously by treating the SNP effects as random effects drawn from a prior distribution. For example, the BayesR model has been implemented for QTL detection in the dairy cattle genome and for human disease traits [15]. The results show that BayesR can increase the power of identifying the known genes in contrast with GBLUP and GWAS.

Even though nonlinear models are attractive, one limitation is that nonlinear models typically require long computational times. Due to the hierarchical estimation of posterior distributions of SNP effects and their variances, nonlinear models have usually been implemented with Markov Chain Monte Carlo (MCMC). This requires a large number of iterations with time per iteration scaled linearly with the number of markers (m) and the number of individuals (n). Genomic data sets are now often very large and are rapidly becoming larger. For human, 300,000 to 9 million SNPs arrays genotyped on up to 253 K individuals [26, 27] are available for association studies and disease/fitness prediction. In dairy cattle, whole genome sequence data including 39 million variants has been published by the 1000 bull genomes project [28]. When confronted with such huge genomic data, Bayesian methods can be so computationally expensive that it is not possible to use them.

Two approaches have been used to develop more computationally efficient algorithms for implementing Bayesian mixture models. One is to modify MCMC with speed-up schemes. For example, Moser et al. [8] introduced a "500SNPs" scheme to pick 500 SNPs with non-zero effects to be updated instead of all the SNPs. Such modification schemes could reduce the running time by 3 ~ 6 fold. Calus et al. [29] proposed a right-hand-side updating algorithm to cluster multiple SNPs (similar to a haplotype) to be updated as one during MCMC iterations. The results on 50 K SNP panels demonstrated up to 90 % reduction in computational time without reducing prediction accuracy. The other approach is to introduce heuristic methods (e.g. iterated conditional expectation, ICE; expectation maximisation, EM) as an alternative to MCMC. There are a wide range of fast versions of Bayesian approaches to genomic prediction using these methods (including fastBayesB, emBayesB, emBayesR) [30–35], which are several orders faster than MCMC implementations. However, none of these algorithms gives consistently as high prediction accuracy as their MCMC counterparts. The EM method of Wang et al. [30], emBayesR, gave higher accuracy than ICE based methods but still had a reduction in accuracy of 5 % ~ 7 % for traits with mutations of moderate to large effect. In other words, the heuristic approximations works best when there are no mutations of moderate to large effect, otherwise accuracy can be compromised. This is undesirable, especially when the largest advantage of the non-linear Bayesian methods over BLUP is observed when there are mutations of moderate to large effect (where moderate effect can mean a QTL explaining 1 % of the variance if the data set is large)!

Motivated by the deficiency of both MCMC (long computing terms) and fast versions of nonlinear models (lower prediction accuracy with some genetic architectures), we hypothesise that a hybrid scheme, beginning with EM iterations and finishing with MCMC sampling iterations, would give similar prediction accuracy to a full MCMC implementation, while having a significant

speed advantage. Here we propose a hybrid algorithm (termed HyB_BR) of Expectation-Maximisation (EM) (emBayesR) and MCMC under the BayesR model. The algorithm also incorporates a speed-up scheme where only a proportion of SNP continue to be sampled in MCMC iterations. In comparison with emBayesR [30], the main improvement is that HyB_BR introduces a limited number of MCMC iterations after EM to improve the solutions from emBayesR.

To evaluate the predictive ability and computational efficiency of HyB_BR, prediction accuracy was compared with BayesR and GBLUP in two data sets. The first data set was 800 K SNP genotypes in 16,214 Holstein and Jersey bulls and cows. The prediction accuracy within these breeds and in a third breed (Aussie Red) was evaluated. The results showed that HyB_BR achieved very similar prediction accuracy to BayesR, while reducing the running time by up to 17 fold, and overcoming the limitations of slightly reduced accuracy of emBayesR. As a result of running the algorithm, the posterior probability of each SNP being in the model was derived, and this was used for QTL mapping. The resulting QTL regions were compared between the approaches and with previous literature reports. The results demonstrated that HyB_BR has enough power to detect the major known genes affecting milk production traits in dairy cattle as well as some novel regions. HyB_BR was also evaluated in a second data set - the Welcome Trust Case Control Consortium (WTCCC) human disease data set [27]. The results demonstrated that HyB_BR is a promising method for risk prediction and genetic architecture inference for human disease traits as well.

Methods

The mixture data model

The overall model at the level of the data for HyB_BR (independent of MCMC and EM implementation) including all the relevant parameters and priors is described first. The model assumes that **y**, the phenotypic records of n individuals, is a linear model of fixed effects (**β**), SNP effect (**g**), random polygenic effects (**v**) and environmental errors (**e**):

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zg} + \mathbf{Wv} + \mathbf{e}, \tag{1}$$

where,

β = vector of p fixed effects, following uninformative priors.

g = vector of m SNP effects. For each SNP, $g_i \sim b(i,1) \times N\left(0, 0 * \sigma_g^2\right) + b(i,2) \times N\left(0, 0.0001 * \sigma_g^2\right) + b(i,3) \times N\left(0, 0.001 * \sigma_g^2\right) + b(i,4) \times N\left(0, 0.01 * \sigma_g^2\right)$, in which σ_g^2 is the genetic variance of the trait and $b(i,k)$ is a scalar with two possible values $\{0,1\}$, determining whether or not the effect of the ith SNP is derived from the kth distribution.

Pr = vector of probabilities where $\boldsymbol{Pr}_k$ = scalar with the value range between 0 and 1. The parameter **Pr** defines the proportion of all the SNPs following each of four normal distributions k. $\boldsymbol{Pr}_k$ is assumed to follow a Dirichlet distribution with the parameter $\alpha = (1, 1, 1, 1)^T$.

v = vector of q polygenic effects (breeding values, for the proportion of variance not explained by the SNP), with $\mathbf{v} \sim N(0, \mathbf{A}\sigma_a^2)$, **A** is the $q \times q$ pedigree-based relationship matrix, σ_a^2 is the polygenic variance, q is the number of individuals in the pedigree.

e = vector of n residual errors. For cattle data, $\mathbf{e} \sim N(0, \mathbf{E}\sigma_e^2)$, where **E** is a $n \times n$ diagonal matrix so that the error variance associated with different records can vary. For example, for bulls, the phenotype would be daughter yield deviations, which would have a lower error variance than the trait deviations (TD) of cows [36]. For human data where all phenotypes have the same magnitude of error, **E** matrix can be replaced by the identity matrix **I**.

X = $n \times p$ design matrix, allocating phenotypes **y** to fixed effects **β**.

Z = $a\ \ n \times m$ genotype matrix with elements $\mathbf{z}_{ij} = \left(\mathbf{s}_{ij} - 2p_i\right) / \sqrt[2]{2p_i(1-p_i)}$, in which $\mathbf{s}_{ij}$ is the genotypes of the jth individual for the ith SNP (0, 1 or 2 copies of the second allele), and p_i is the allele frequency of each SNP i.

W = $n \times q$ design matrix, aims at allocating the $q \times 1$ vector of polygenic effects to **y**.

Note that model (1) extends the model used by Wang et al. [30] to include fixed effects, polygenic effects and weights.

The prior distribution of each SNP effect g_i conditional on $b(i,k)$ is

$$\mathrm{p}(g_i|b(i,k)) = \begin{cases} \delta(g_i),\ b(i,1) = 1 \\ \frac{1}{\sqrt{2\pi\sigma_i^2[k]}} \exp\left(-\frac{g_i^2}{2\sigma_i^2[k]}\right),\ b(i,k) = 1 (k = 2,3,4) \end{cases}.$$

Where, $\delta(g_i)$ denotes the Dirac delta function with all probability mass at $g_i = 0$.

The joint distribution $\mathrm{p}(g_i, b(i,k)|Pr_k)$ (i.e. conditional on Pr_k) can be written as:

$$\begin{aligned} p(g_i, b(i,k)|Pr_k) &= \prod_{k=1}^{4} p\left(g_i|b(i,k)\right) \times p(b(i,k)|Pr_k) \\ &= \left(\delta(g_i)Pr_1\right)^{b(i,1)} \prod_{k=2}^{4} \left(\frac{1}{\sqrt{2\pi\sigma_i^2[k]}} \exp\left(-\frac{g_i^2}{2\sigma_i^2[k]}\right) Pr_k\right)^{b(i,k)} \end{aligned} \tag{2}$$

The implementation of HyB_BR with the mixture model defined above consists of two components: 1) The Expectation-Maximization module. HyB_BR first implements the EM iterations under the mixture Gaussian model (Eq. 2) to give approximations for the

parameter set including SNP effects **g**, proportion of SNP in each distribution **Pr**, error variance σ_e^2, and polygenic variance σ_a^2. For the estimation of each SNP effect, the PEV (predicted error variance) correction is introduced to account for the errors which are generated from the estimations of all other SNP effects (detailed in Additional file 1). 2) MCMC module. Once the EM steps are converged, the output values of the parameters are used in the modified MCMC iterations as the start values. For the final step, a MCMC scheme is implemented with a limited number of iterations.

EM module

In the following EM modules, the parameter set $\theta = \{\mathbf{g}, \mathbf{Pr}, \boldsymbol{\beta}, \mathbf{v}, \sigma_e^2\}$ will be estimated by their maximum a posteriori (MAP) value. Similar to emBayesR [30], all the parameters θ were estimated according to the expectation-maximisation process with steps:

i) Define the log likelihood $f(\mathrm{y}|\theta)$ of the data under the data model (1).
ii) Derive the log posterior function of the parameters using Bayes' theorem. Following Bayes' theorem, the log posterior distribution of the parameter sets θ is based on the rule $logp(\theta|\mathrm{y}) \propto logf(\mathrm{y}|\theta) + logp(\theta)$, with $p(\theta)$ the prior for the parameter.
iii) Take the expectation on the posterior function over the missing data.
iv) Differentiate the expected posterior function and solve for this equal to zero to obtain MAP (Maximum A Posterior) of the parameter set θ.

In the Expectation-maximization implementation, the posterior distributions for each parameter $p(\theta|\mathbf{y})$ are obtained while "integrating out" the other parameters. For example, for the estimation of each SNP effect g_i (SNP i in the vector **g**), we maximize the posterior distribution of each SNP effect $p(g_i|\mathbf{y}, b(i,k), \mathrm{Pr_k}, \boldsymbol{\beta}, \mathbf{v}, \sigma_e^2)$ by integrating out the other SNP effects $g_{j \neq i}$, the parameters $b(i,k)$, $\boldsymbol{\beta}$, **v**, but we fix the proportion parameter $\mathrm{Pr_k}$ and the error variance σ_e^2 at their maximum posterior estimates. In the following, we will detail the inference process for several key parameters including SNP effects (**g**), the mixing parameters (Pr_k), fixed effects ($\boldsymbol{\beta}$), polygenic effects (**v**) and the error variance (σ_e^2) separately:

1) Estimation of SNP effects g

As in our EM version of BayesR [30], in HyB_BR we will update the estimated effect of SNPs one at a time. Therefore, we rewrite **Zg** in Eq. (1) as the sum of the effect of the current SNP $\mathbf{Z_i}g_i$ and the combined effect of all other SNP effects $\mathbf{u}_i$ ($\mathbf{u}_i = \sum_{j \neq i}\mathbf{Z_j}g_j$). We rewrite the model (1) as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z_i}g_i + \mathbf{u}_i + \mathbf{W}\mathbf{v} + \mathbf{e} \tag{3}$$

where, g_i(the effect of SNP i) is the ith element of the vector **g**, and $\mathbf{u}_i = \sum_{j \neq i}\mathbf{Z_j}g_j$.

We estimate of $\hat{g}_i$ by the value of g_i that maximises the posterior probability $\mathrm{P}\left(g_i|\mathbf{y}, \hat{\mathbf{P}}\mathbf{r}, \widehat{\sigma_e^2}\right)$ where $\hat{\mathbf{P}}\mathbf{r}$ and $\widehat{\sigma_e^2}$ are the MAP estimates of **Pr** and σ_e^2 conditional on **y**.

To perform this, we first introduce "missing data" ($b(i,k)$, $\boldsymbol{\beta}$, **v**, $\mathbf{u}_i$), and then "integrate them out" via the Expectation-Maximisation steps. In detail, the marginal posterior distribution of each SNP effect g_i can be written as:

$$p\left(g_i, b(i,k)|\mathbf{y}, \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}_i, \widehat{\sigma_e^2}, \widehat{Pr}_k\right) \propto p\left(\mathbf{y}|g_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}_i, \widehat{\sigma_e^2}, \widehat{Pr}_k\right) p\left(g_i, b(i,k)|\widehat{Pr}_k\right).$$

Under the model (3), the first term $p\left(\mathbf{y}|g_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}_i, \widehat{\sigma_e^2}, \widehat{Pr}_k\right)$ is obtained according to the following normal distribution:

$$\mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathbf{Z_i}g_i - \mathbf{W}\mathbf{v} - \mathbf{u}_i \sim N(0, \mathbf{E}\sigma_e^2),$$

which can be transformed as:

$$\mathbf{e}^* - \mathbf{Z_i}g_i \sim N\left(0, \mathbf{E}\sigma_e^2\right),$$

Where, $\mathbf{e}^* = \mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{v} - \mathbf{u}_i$.

Therefore, the term $p\left(\mathbf{y}|g_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}_i, \widehat{\sigma_e^2}, \widehat{Pr}_k\right)$ can be written as:

$$p\left(\mathbf{y}|g_i, \mathbf{u}_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \widehat{\sigma_e^2}, \widehat{Pr}_k\right) = \frac{1}{\left(2\pi\widehat{\sigma_e^2}\right)^{\frac{n}{2}}}\frac{1}{|\mathbf{E}|}exp\left[-\frac{1}{2\widehat{\sigma_e^2}}\left(\mathbf{e}^* - \mathbf{Z_i}g_i\right)'\mathbf{E}^{-1}\left(\mathbf{e}^* - \mathbf{Z_i}g_i\right)\right]$$

Then the log likelihood function is:

$$logp\left(\mathbf{y}|g_i, \mathbf{u}_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \widehat{\sigma_e^2}, \widehat{Pr}_k\right) = -\frac{\mathrm{n}}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| - \frac{1}{2\widehat{\sigma_e^2}}\left(\mathbf{e}^* - \mathbf{Z_i}g_i\right)'\mathbf{E}^{-1}\left(\mathbf{e}^* - \mathbf{Z_i}g_i\right) \tag{4}$$

Ignoring an additive constant, the second term $p\left(g_i, b(i,k)|\widehat{Pr}_k\right)$ is defined in the Eq. (2). Then the log of $p\left(g_i, b(i,k)|\widehat{Pr}_k\right)$ is:

$$logp\left(g_i, b(i,k)|\widehat{Pr}_k\right) = b(i,1)\log\left(\delta\left(g_i\right)\widehat{Pr}_1\right) + \sum_{k=2}^{4} b(i,k)\left(-\frac{1}{2}\log\sigma_i^2[k] - \frac{g_i^2}{2\sigma_i^2[k]} + \log\widehat{Pr}_k\right) \quad (5)$$

Treating $(\mathbf{e}^*, b(i,k))$ as missing data and omitting the terms without g_i, the expectation of the log marginal posterior of each SNP effect is:

$$E_{\mathbf{e}^*, b(i,k)}\, logp\left(g_i, b(i,k)|\mathbf{y}, \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}_i, \widehat{\sigma_e^2}, \widehat{Pr}_k\right) = E_{\mathbf{e}^*, b(i,k)} logp\left(\mathbf{y}|g_i, \mathbf{u}_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \widehat{\sigma_e^2}, \widehat{Pr}_k\right) + E_{\mathbf{e}^*, b(i,k)} logp\left(g_i, b(i,k)|\widehat{Pr}_k\right)$$

According to Eq (4), the expectation of the first term is:

$$E_{\mathbf{e}^*, b(i,k)}\, logp\left(\mathbf{y}|g_i, \mathbf{u}_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \widehat{\sigma_e^2}, \widehat{Pr}_k\right) \propto -\frac{1}{2\widehat{\sigma_e^2}}\left\{(\mathbf{e}^* - \mathbf{Z}_i g_i)'\mathbf{E}^{-1}(\mathbf{e}^* - \mathbf{Z}_i g_i) + tr\left(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e}^*)\right)\right\} \quad (6)$$

According to the Eq. (5), the expectation of the second term is:

$$E_{\mathbf{e}^*}, b(i,k)\, logp\left(g_i, b(i,k)|\widehat{Pr}_k\right) \propto P(i,1)\log\left(\delta\left(g_i\right)\widehat{Pr}_1\right) + \sum_{k=2}^{4} P(i,k)\left(-\frac{1}{2}\log\sigma_i^2[k] - \frac{g_i^2}{2\sigma_i^2[k]} + \log\widehat{Pr}_k\right) \quad (7)$$

Where, $P(i,k) = E\left(b(i,k)|\mathbf{y}, \widehat{Pr}_k\right)$. The term $P(i,k)$ can be derived as in the Additional file 2.

Hence, in the Maximisation steps of EM, we differentiate Eqs. (6) and (7) with respect to $\hat{g}_i$, and then obtain an estimate for the SNP effect as:

$$\frac{\partial E_{\mathbf{e}^*}, b(i,k) logp\left(g_i, \mathbf{u}_i, b(i,k)|\mathbf{y}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{v}}, \widehat{\sigma_e^2}, \widehat{Pr}_k\right)}{\partial g_i} = \left[-\sum_{k=2}^{4}\frac{P(i,k)}{\sigma_i^2[k]} - \frac{\mathbf{Z}_i'\mathbf{E}^{-1}\mathbf{Z}_i}{\widehat{\sigma_e^2}}\right]g_i + \frac{\mathbf{Z}'\mathbf{E}^{-1}\mathbf{e}^*}{\widehat{\sigma_e^2}} = 0$$

$$\hat{g}_i = \left[\mathbf{Z}_i'\mathbf{E}^{-1}\mathbf{Z}_i + \sum_{k=1}^{4}\left(P(i,k)\frac{\widehat{\sigma_e^2}}{\sigma_i^2[k]}\right)\right]^{-1}\left[\mathbf{Z}'\mathbf{E}^{-1}\mathbf{e}^*\right] \quad (8)$$

2) Estimation of parameter Pr

This follows a common method for an EM algorithm to analyse a mixture of distributions. We introduce the 'missing data' $b(i,k)$ which is the indicator variable that indicates which of the $k = 4$ distributions SNP effect i is drawn from. The posterior distribution of proportion parameter **Pr** is:

$$p(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b})p(\mathbf{b}|\mathbf{Pr})p(\mathbf{Pr})$$

Where,

The term $p(\mathbf{y}|\mathbf{b})$ does not involve **Pr**. So when we differentiate with respect to **Pr**, this term will drop out and therefore can be ignored.

$$p(\mathbf{b}|\mathbf{Pr}) = \prod_{i=1}^{n}\prod_{k=1}^{4}(Pr_k)^{b(i,k)}$$

$$p(\mathbf{Pr}) = \prod_{k=1}^{4} Pr_k$$

Therefore, the log posterior expression of **Pr** can be written as:

$$logp(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) \propto logp(\mathbf{b}|\mathbf{Pr}) + logp(\mathbf{Pr}) = \sum_{i=1}^{n}\sum_{k=1}^{4} b(i,k) logPr_k + \sum_{k=1}^{4} logPr_k.$$

Treating **b** as the missing data and defining $P(i,k) = E(b(i,k)|\mathbf{y}, Pr_k)$, the expectation of the posterior can be written as:

$$E_{\mathbf{b}|\mathbf{y}} logp(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) = \sum_{i=1}^{n}\sum_{k=1}^{4} P(i,k) logPr_k + \sum_{k=1}^{4} logPr_k. \quad (9)$$

Introducing Lagrange multiplier λ to account for the constraint that $\sum_{k=1}^{4} Pr_k = 1$ and differentiate with respect to Pr_k, the parameter **Pr** can be estimated by:

$$\frac{\partial E_{\mathbf{b}|\mathbf{y}}\left[logp(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) + \lambda\left(\sum_{k=1}^{4} Pr_k - 1\right)\right]}{\partial Pr_k} = \frac{\sum_{i=1}^{m} P(i,k)}{Pr_k} + \frac{1}{Pr_k} + \lambda = 0$$

$$Pr_k = \frac{\sum_{i=1}^{m} P(i,k) + 1}{\sum_{k=1}^{4}\left(\sum_{i=1}^{m} P(i,k) + 1\right)} \quad (10)$$

The computation of $P(i,k)$ is given in the Additional file 2.

3) Estimation of fixed effects (β) and the error variance (σ_e^2)

Since fixed effects (**β**) and the error variance has uninformative priors, their posterior distribution is:

$$p(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y}) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}}\frac{1}{|\mathbf{E}|}exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\hat{\mathbf{v}})'\mathbf{E}^{-1}(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\hat{\mathbf{v}})\right]$$

As $\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\hat{\mathbf{v}} = \mathbf{e}$, the full log likelihood based on this model is:

$$logp(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y}) = -\frac{n}{2}\log\sigma_e^2 + \frac{1}{2\sigma_e^2}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e} \quad (11)$$

Treating $\mathbf{e}$ as the missing data, the expectation of the Eq. (11) can be expressed as

$$\begin{aligned} E_{\mathbf{e}|\mathbf{y}}\, logp(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y}) &= E_{\mathbf{e}|\mathbf{y}}\left[-\frac{n}{2}\log\sigma_e^2 + \frac{1}{2\sigma_e^2}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e}\right] \\ &= -\frac{n}{2}\log\sigma_e^2 + \frac{1}{2\sigma_e^2}\left[\mathbf{e}'\mathbf{E}^{-1}\mathbf{e} + tr\left(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e})\right)\right] \end{aligned}$$

In theory, PEV($\mathbf{e}$) ≠ PEV($\mathbf{e}^*$) due to $\mathbf{e} = \mathbf{e}^* + \mathbf{Z}_i g_i$. However, since each SNP effect is shrunk severely towards zero by GBLUP [4], we approximate PEV($\mathbf{e}$) ≅ PEV($\mathbf{e}^*$). The calculation of the term PEV($\mathbf{e}^*$) is detailed in the Additional file 1.

Therefore, differentiating the equation $E_{\mathbf{e}|\mathbf{y}} logp(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y})$ with regard to σ_e^2 and $\mathbf{b}$, we achieve:

$$\widehat{\sigma_e^2} = \frac{1}{n}\left[(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\hat{\mathbf{v}})'\mathbf{E}^{-1}(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\hat{\mathbf{v}}) + tr\left(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e}^*)\right)\right] \quad (12)$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{E}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{E}^{-1}(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{W}\hat{\mathbf{v}}) \quad (13)$$

4) Estimation of polygenic effects (v)

Under the model (1), the conditional posterior density function of polygenic effects $\mathbf{v}$ is:

$$p(\mathbf{v}|\mathbf{y}) = p\left(\mathbf{y}|\mathbf{v}, \hat{\mathbf{g}}, \hat{\boldsymbol{\beta}}, \widehat{\sigma_e^2}\right)p(\mathbf{v})$$

Where,

$$p\left(\mathbf{y}|\mathbf{v}, \hat{\mathbf{g}}, \hat{\boldsymbol{\beta}}, \widehat{\sigma_e^2}\right) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}}\frac{1}{|\mathbf{E}|}exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\hat{\mathbf{g}}-\mathbf{W}\mathbf{v})'\mathbf{E}^{-1}\left(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{W}\mathbf{v}\right)\right] \quad (14)$$

$$p(\mathbf{v}) = \frac{1}{(2\pi\sigma_a^2)^{\frac{q}{2}}}\frac{1}{|\mathbf{A}|}exp\left[-\frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \quad (15)$$

Therefore, the log posterior based on Eqs. (14) and (15) is:

$$\begin{aligned} logp(\mathbf{v}|\mathbf{y}) &= logf\left(\mathbf{y}|\mathbf{v}, \hat{\mathbf{g}}, \hat{\boldsymbol{\beta}}, \widehat{\sigma_e^2}\right) + logp(\mathbf{v}) \\ &= \left[-\frac{n}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| + \frac{1}{2\widehat{\sigma_e^2}}\left(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{W}\mathbf{v}\right)'\mathbf{E}^{-1}\left(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{W}\mathbf{v}\right)\right] \\ &\quad + \left[-\frac{q}{2}\log\sigma_a^2 - \log|\mathbf{A}| + \frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \end{aligned} \quad (16)$$

According to the equation $\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\hat{\mathbf{v}} = \mathbf{e}$, the Eq. (16) can be written as:

$$\begin{aligned} logp(\mathbf{v}|\mathbf{y}) &= logf\left(\mathbf{y}|\mathbf{v}, \hat{\mathbf{g}}, \hat{\boldsymbol{\beta}}, \widehat{\sigma_e^2}\right) + logp(\mathbf{v}) \\ &= \left[-\frac{n}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| + \frac{1}{2\widehat{\sigma_e^2}}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e}\right] \\ &\quad + \left[-\frac{q}{2}\log\sigma_a^2 - \log|\mathbf{A}| + \frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \end{aligned} \quad (17)$$

Taking expectation over the missing data $\mathbf{e}$, we get:

$$\begin{aligned} E_{\mathbf{e}|\mathbf{y}} logp(\mathbf{v}|\mathbf{y}) &= \left[-\frac{n}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| + \frac{1}{2\widehat{\sigma_e^2}}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e} + tr\left(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e})\right)\right] \\ &\quad + \left[-\frac{q}{2}\log\sigma_a^2 - \log|\mathbf{A}| + \frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \end{aligned} \quad (18)$$

Differentiating the Eq. (18) with regards to $\mathbf{v}$, we get:

$$\hat{\mathbf{v}} = \left(\mathbf{W}'\mathbf{E}^{-1}\mathbf{W}\sigma_a^2 + \sigma_e^2\mathbf{A}^{-1}\right)^{-1}\sigma_a^2\mathbf{W}'\mathbf{E}^{-1}\left(\mathbf{y}-\mathbf{Z}\hat{\mathbf{g}}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) \quad (19)$$

Where, for simplicity, the variance σ_a^2. will be fixed as the specified value from GBLUP estimation.

Table 1 lists all the parameters and their equation derived from EM steps.

Steps for EM module

The overall procedure of EM is described by means of the pseudo code, steps ①~⑦. Here we will detail these steps according to their sequence appearing in the pseudocode descriptions:

Step *EM*_①: Initialise the parameters g, Pr, σ_i^2 and Construct X, A, G, E, W matrices. Similar to emBayesR [30], the starting values of g and Pr were set as g = 0.01 and Pr = {0.5, 0.487, 0.01, 0.003}, while σ_i^2 = {0, 0.0001 * σ_g^2, 0.001 * σ_g^2, 0.01 * σ_g^2}. The genetic variance σ_g^2 and polygenic variance σ_a^2 are obtained

Table 1 The list of all the estimated parameters including the possibility for each SNP (**P**(**i**, **k**)), the proportion parameter (**Pr**), each SNP effect ($\mathbf{g_i}$), error variance ($\boldsymbol{\sigma_e^2}$), fixed effect (**β**), and polygenic effects **v** and the according equation derived from EM steps

Parameters	The data model	According equations derived from EM
$E_{e^*} logP(i,k)$	The expected likelihood parameters for $P(i,k)$	Equation (S3)
$P(i,k)$	SNP effects related parameters under the extended model (3)	Equation (S4)
Pr		Equation (10)
g_i		Equation (8)
σ_e^2	The overall model (1)	Equation (12)
β		Equation (13)
v		Equation (19)

from GBLUP. Both variances won't be updated during EM iterations.

The $n \times 3$ matrix X is a design matrix, allocating the phenotypes to fixed effects. In our case, the matrix X is set up with the first column being the mean, the second and third columns defining the breeds (Holstein or Jersey) and sex (bulls or cows) of the cattle. For example, if the ith animal is a Holstein bull, then $x_{i,2} = 1$ with $x_{i,3} = 0$.

The Pedigree relationship matrix A is built using Henderson's rules [37]; while the genomic relationship matrix G is constructed using the equation $G = ZZ'/n$. Diagonal error matrix E is calculated following Garrick et al. [36], and the index matrix W maps individuals in the pedigree to the phenotypes if they have ones.

Step *EM*_②: Calculate the PEV matrix under model 1 (Additional file 1). Then using PEV matrix, calculate $tr\left(\mathbf{E}^{-1}\mathbf{Z_i}\mathbf{Z_i'}\mathbf{E}^{-1}\mathbf{PEV_{u_1}}(\mathbf{e})\right)$ which is used in the equation for $E_{e^*} logP(i,k)$ (Additional file 2). In theory, the calculation of this term should be updated each EM iteration, which is time consuming. To avoid huge computational burden, the PEV matrix is treated as constant value for the term $tr\left(E^{-1}Z_iZ_i'E^{-1}PEV_{u_1}(e)\right)$ in front of EM loop.

Then for each SNP i (i in 1 to m)

Step *EM*_③: Correct y for the effects of all other SNPs except current SNP i with equation $\mathbf{y}^{\dagger} = \mathbf{y} - \sum_{j \neq i} \mathbf{Z_j}\ \hat{g}_j - \mathbf{X}\hat{\mathbf{b}} - \mathbf{W}\hat{\mathbf{v}}$.

Step *EM*_④: Estimate the probability that the effect of SNP i is from one of four normal distributions $E_{e^*} logP(i,k)$ with the equation (S3). After this, $P(i,k)$ is calculated with the equation $exp\left(E_{e^*} logP_{ik} / \sum_{k=1}^{4} exp\ \left(E_{e^*} logP_{ik}\right)\right)$ (S4).

Step *EM*_⑤: the SNP effect $\hat{g}_i$ is updated via Eq. (8). After effects have been estimated for all SNP,

Step *EM*_⑥: Estimate σ_e^2 with Eq. (12), fixed effects β with Eq. (13), update Pr_k with Eq. (10), and update polygenic effects v with the Eq. (19).

Step *EM*_⑦: Assess convergence criterion $(\hat{g}^l - \hat{g}^{l-1})'(\hat{g}^l - \hat{g}^{q-1})/((\hat{g}^{l\prime}\hat{g}^l) \leq 10^{-10}$ with l being the EM iterations number. If not converged, then return to Step ③ for the next EM iteration; otherwise, exit the EM iterations and return the estimates of parameters from the final iterations.

Modified MCMC module with speed-up scheme

The outputs of the EM including SNP solutions, polygenic effects, error variance and genetic variance are used as starting values of parameters for the MCMC module, which allows MCMC to begin with no burn in.

The MCMC module of HyB_BR implements the same Gibbs sampling processes as BayesR [15] but modified with one speed-up scheme as follows. Over the first 500 iterations, the average probability that the SNP effect is zero ($p(i, 1)$) is calculated. If $p(i, 1) \geq a$, then the SNP effect is set to zero and is not updated in future iterations.

The test for selecting a reasonable value of the parameter a was conducted as follows. The impact of value of a from 0.85 to 0.95 on prediction accuracy was investigated for the milk production traits and fertility, Fig. 1. The results show that criterion $p(i, 1) \geq, 1$, is the lowest threshold which gives an accuracy very close to the maximum. The criterion means SNP i has more than 90 % probability of having no effect.

The modified MCMC steps can then be described as follows:

Step *MCMC*_①: sampling the error variance $\hat{\sigma}_e^2$ according to the distribution $\hat{\sigma}_e^2 \sim Inv\text{-}\chi^2\left(\frac{n-2, y^{*\prime}E^{-1}y^*}{n-2}\right)$, with $y^* = \left(y - Zg - X\hat{\beta} - W\hat{v}\right)$.

Step *MCMC*_②: sampling the fixed effects β from the distribution $N\left(\beta_\mu, \left(X'E^{-1}X\right)^{-1}\hat{\sigma}_e^2\right)$, with $\beta_\mu = \left(\mathbf{X'E^{-1}X}\right)^{-1}\mathbf{X'E^{-1}}(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{W}\hat{\mathbf{v}})$.

Step *MCMC*_③: Polygenic variance is sampled $\hat{\sigma}_a^2 \sim Inv\text{-}\chi^2\left(\frac{n-2, \hat{\mathbf{v}}'\mathbf{A}^{-1}\hat{\mathbf{v}}}{n-2}\right)$.

Step *MCMC*_④: The polygenic effects are sampled from normal distribution $N(\mu, s)$, with the mean $\mu = \hat{v}$ from Eq. (19) and the variance $s = \left(W'E^{-1}W + A^{-1}\sigma_e^2/\sigma_a^2\right)^{-1}$.

Then for each SNP i (i in 1 to m),

Step *MCMC*_⑤: Implement the speed-up scheme : if (iterations > 500) and ($P(i, 1) > 0.9$), then stop updating this SNP i.

Else,

Step *MCMC*_⑥: Estimate the log likelihood that the effect of SNP i is from one of four normal distributions $L(g_i|\sigma_i^2[k])$. Following the derivation steps of Kemper et al. [15], the optimised equation of the log likelihood function is

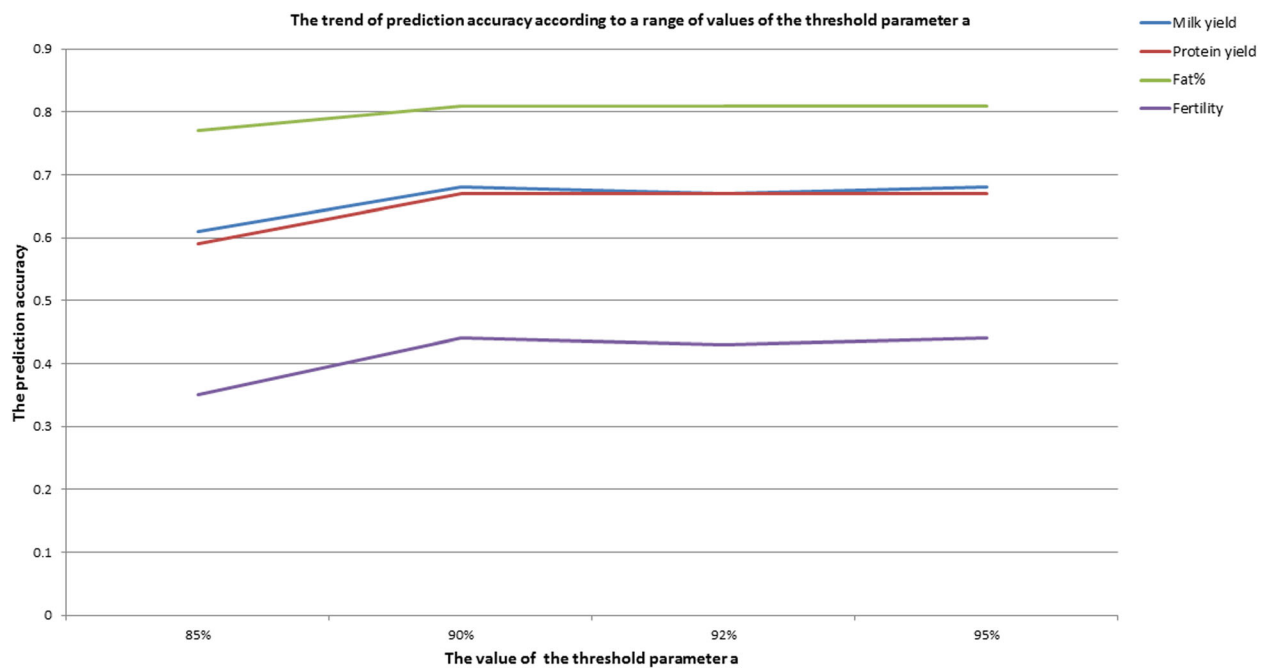

Fig. 1 The trend of prediction accuracy according to a range of values of the threshold parameter **a**

$$L(g_i|\sigma_i^2[k]) = -\frac{1}{2}\left\{\log\left(\sigma_i^2[k]Z_i'Z_i+\sigma_e^2\right)+\left((e^*)'E^{-1}Z_i\right)^2\sigma_i^2[k]\sigma_{e^{-2}}/\left(\sigma_i^2[k]Z_i'E^{-1}Z_i+\sigma_e^2\right)\right\} + log\,\mathrm{Pr}_k,$$

with $e^* = y - X\beta - u - Wv$.

After this, $P(i, k)$ is calculated with the equation:

$$exp\left(L(g_i|\sigma_i^2[k])\right)/\sum_{k=1}^{4} exp\ \left(L(g_i|\sigma_i^2[k])\right)$$

Step *MCMC*_⑦: Sample $\hat{g}_i$ with $N(\mu, s)$,

$$\mu = \left[\mathbf{Z}_i'\mathbf{E}^{-1}\mathbf{Z}_i + \frac{\hat{\sigma}_e^2}{\sigma_i^2[k]}\right]^{-1}\left[\mathbf{Z}'\mathbf{E}^{-1}\,e^*\right], \text{ and } s = \left[Z_i'E^{-1}Z_i + \frac{\hat{\sigma}_e^2}{\sigma_i^2[k]}\right]^{-1}.$$

Step *MCMC*_⑧: Update Pr ~ Dirichlet($\beta_1 + 1, \beta_2 + 1, \beta_3 + 1, \beta_4 + 1$),where $\beta_1, \beta_2, \ldots, \beta_4$ are the number of SNPs in one of four normal distributions.

Return to MCMC step 1.

HyB_BR was written in the C++ programming language.

Data

Cattle

One thousand seven hundred forty-five Holstein and Jersey cattle and 114 Australian Red bulls were genotyped with the 777 K Illumina HD bovine SNP chip. 15,049 Holstein and Jersey bulls and cows, 249 Australian red bulls and cows were genotyped with the 54 K Illumina Bovine SNP array. After stringent quality control and SNP filtering described in [14], there were 632,003 SNPs remaining for animals genotyped with the 777 K SNP panel, and 43,425 SNPs remaining for animals genotyped with the 54 K SNP array. Animals genotyped with the 43,425 SNPs, were imputed to 632,003 SNP genotypes using Beagle 3.0 [38]. Therefore, the total data set was 17,157 cattle of three breeds with real or imputed genotypes for 632,003 SNP.

The phenotypes include milk yield, protein yield, fat percent(fat%), and cow fertility. The heritability of

Table 2 The number of individuals in the reference sets and validations sets related to three traits including Milk yield (MilkY), Protein yield (ProtY), Fat Percent (Fat%) and Fertility

Traits	Reference Sets				Validation Sets		
	Holstein		Jersey		Holstein Bulls	Jersey Bulls	Australian Red Bulls
	Bulls	Cows	Bulls	Cows			
MilkY/ProtY/Fat%	3049	8478	770	3917	262	105	114
Fertility	2806	7838	716	3830	396	81	114

Table 3 Three input variance parameters related to the reference data sets

Reference Set	Traits	σ_e^2	σ_g^2	σ_v^2
Holstein and Jersey bulls & cows	Milk yield	133284.0	108619.0	34925.6
	Protein yield	132.579	68.6635	29.1662
	Fat%	0.0180012	0.0575729	0.0127094
	Fertility	3283.80	31.6187	0.000332297

The variances including error variance ($\boldsymbol{\sigma_e^2}$), genetic variance ($\boldsymbol{\sigma_g^2}$), and polygenic variance ($\boldsymbol{\sigma_v^2}$) are estimated by ASReml 4

these traits varies from 0.33 (for milk yield, protein yield and fat%), to 0.03 (for cow fertility). The fertility (reproductive performance of dairy cows) is usually measured according to calving interval (CI, the number of days between successive calvings), days from calving to first service (CFS), pregnancy diagnosis, lactation length (LL), and survival to second lactation on Australian Holstein and Jersey cows [39, 40]. Here, the fertility phenotype was calving interval (CI). Here, the fertility phenotype is mainly derived from CI. The phenotypes for all the traits were daughter trait deviations (DTD) for bulls (the average of their daughters phenotypes, corrected for fixed effects), and trait deviations (TD) for cows (as described by Kemper et al. [15]). For genomic prediction, the data was separated into a reference set, where SNP effects were estimated, and validation sets, where the prediction accuracy was assessed, and the division of animals into reference and validation sets was by year of birth (youngest animals in validation set). The reference data includes bulls and cows from two breeds (Holstein and Jersey), and the predictions were evaluated in the other animals of the same breeds or in a breed (Aussie red) not included in the reference set. The exact number of individuals in these data sets for each trait is given in Table 2.

To compare the computational time required by the different genomic prediction methods, we also used three reference sets with increasing different numbers of animals; Ref 1_ CATTLE had 3049 Holstein bulls; Ref 2_CATTLE had 11,527 Holstein bulls and cows, while Ref 3_CATTLE was the complete reference data set with 16,214 animals.

For the EM module, estimates of three variance components (σ_e^2, σ_v^2, σ_g^2) were required as the input. We ran Asreml4.0 [41] (which is implemented with GBLUP methods) on these data sets to estimate these variance parameters and the results are listed in Table 3.

The correlation between GEBV and DTD in the validation sets was used as a proxy for prediction accuracy. The regression of DTD on GEBV in the validation sets was used to investigate if any of the methods resulted in biased predictions.

Case/Control human disease trait data

For predicting human disease risk, seven disease traits of the Welcome Trust Case Control Consortium (WTCCC) genomic data [27] including bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), Hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D) were used. Following the steps of strict QC on SNP data [7, 8, 42] with Plink [43], we had seven combined case/control data sets (one for each trait) with different number of markers and records listed in Table 4. The input parameters of seven datasets for HyB_BR including error variance and genetic variance were calculated by GCTA [44], given in Table 4. To assess prediction accuracy, for each disease, we randomly generated 20 splits of the data with 80 % of individuals for the reference set and 20 % for the validation set. To assess the prediction ability, the area under the ROC curve (AUC) [45] was calculated.

Results

Compute time comparison of HyB_BR and BayesR

To compare computational efficiency, HyB_BR without the speed-up scheme (labelled as Hyb_BR_Orig), HyB_BR with the speed-up scheme and pure MCMC BayesR were implemented on three data sets with 632,003 markers but different numbers of records, varying from 3049 in Ref 1_CATTLE, 11,527 in Ref 2_CATTLE, to 16,214 in Ref 3_CATTLE. As used by Kemper et al. [15], pure MCMC BayesR required 40,000 iterations of complexity $O(mn)$

Table 4 The size and genetic architecture of seven combined control/case data sets

Diseases	Number of records	Number of markers	σ_e^2	σ_g^2	h^2
BD	4722	292,496	0.070509	0.17156	0.71
CAD	4864	296,610	0.149782	0.09189	0.38
CD	4577	301,579	0.073900	0.16056	0.69
HT	4890	294,404	0.113621	0.12816	0.53
RA	4704	295,890	0.070900	0.07120	0.50
T1D	4824	296,228	0.064739	0.12567	0.66
T2D	4722	294,641	0.099866	0.14497	0.59

The error variance (σ_e^2) and genetic variance (σ_g^2) are estimated by GCTA; the heritability (h^2) is estimated by the equation $h^2 = \frac{\sigma_g^2}{(\sigma_e^2+\sigma_g^2)}$

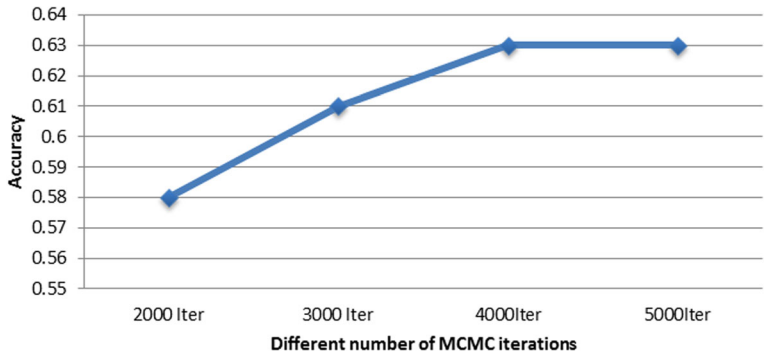

Fig. 2 Prediction accuracy with an increasing number of MCMC iterations for BayesR

with parameters estimated from samples from the posterior distributions (*m* is the number of markers and *n* is the number of individuals). The first 20,000 iterations were removed as burn in. The MCMC module of HyB_BR used only 4000 iterations and burn-in was replaced by the EM (400 iterations to convergence). 4000 cycles for the MCMC module were used after comparing results with increasing number of iterations. The results showed that 4000 were necessary to achieve maximum prediction accuracy (Fig. 2).

The prediction accuracy was evaluated for milk yield with a reference set containing the Holstein and Jersey bulls&cows data.

With the smallest data set (Ref 1_CATTLE), 5 h compute time was required for HyB_BR compared with 96 h for BayesR MCMC (Fig. 3); 35 h required by HyB_BR instead of 410 h of BayesR for Ref 2_CATTLE; And in Ref 3_CATTLE, 42 h for HyB_BR_sp but 720 h for BayesR. These results suggest HyB_BR is at least 10 times faster than BayesR MCMC, with the speed advantage increasing as data sets became larger (17 times faster with the largest data set). The HyB_BR speed up scheme reduced compute time by approximately 50 %, compared to HyB_BR_Orig without the speed up scheme (Fig. 3), with no reduction in the prediction accuracy (Tables 5, 6 and 7).

These timings were recorded on a server with Intel E5-2680 2.7GHz processors and 384GB of 1333 MHz RAM.

The accuracy and bias of within-breeds, multi-breeds and across-breeds prediction for four complex dairy traits

Genomic prediction with a single breed reference

For the within-breed prediction (that is, when a Holstein reference was used to estimate SNP effects used for calculating GEBV in a Holstein validation set, and likewise for Jersey) in Table 5, HyB_BR performed as well as BayesR for all traits, including fat%. Both

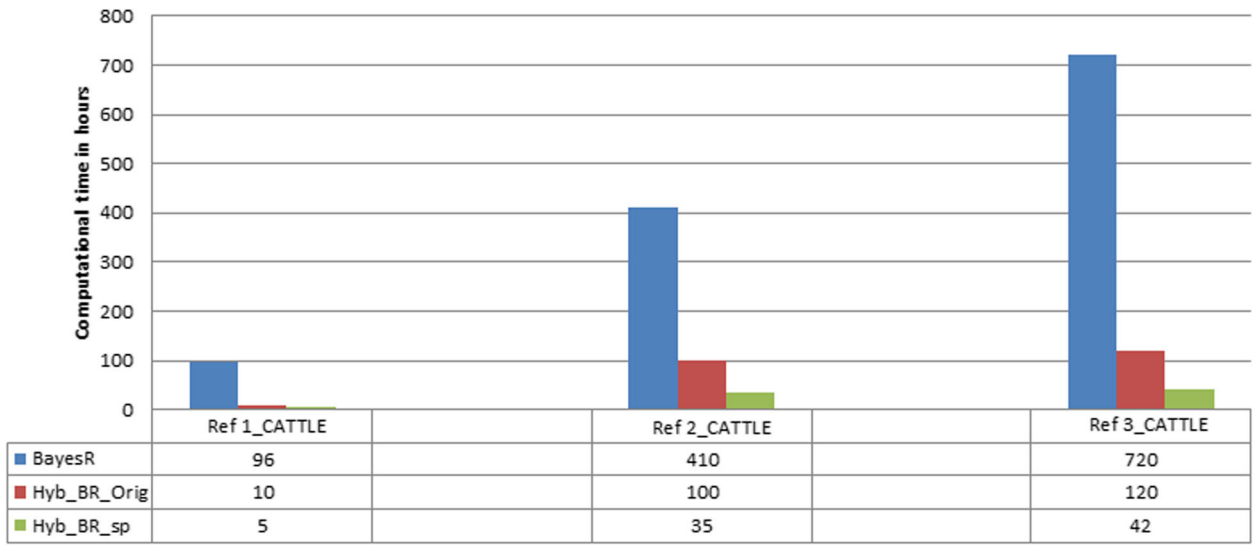

Fig. 3 Computational time in hours required for BayesR, HyB_BR_Orig, and HyB_BR_sp on three reference sets (Ref 1_CATTLE, Ref 2_CATTLE, Ref 3_CATTLE)

Table 5 The accuracy and bias of with-breed prediction of GBLUP, BayesR(BR), emBayesR (EM), and HyB_BR (HB)

		Milk yield		Protein yield		Fat%		Fertility	
		Acc.	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
		Holstein reference to predict Holstein validation							
GBLUP	+Poly[a]	0.57	0.96	0.63	0.98	0.73	0.96	0.43	1.26
	-Poly[b]	0.56	0.86	0.59	0.87	0.71	1.15	0.42	1.27
BR	+Poly[a]	0.63	0.91	0.64	1.01	0.79	1.06	0.43	1.19
	-Poly[b]	0.61	1.00	0.63	1.06	0.77	1.13	0.41	1.19
EM	+Poly[a]	0.62	0.79	0.63	0.85	0.77	0.98	0.42	1.15
	-Poly[b]	0.62	0.92	0.62	0.94	0.74	1.06	0.41	1.15
HB	+Poly[a]	0.63	0.93	0.63	0.97	0.79	1.09	0.43	1.19
	-Poly[b]	0.63	1.03	0.62	1.06	0.76	1.17	0.42	1.19
		Jersey reference to predict Jersey validation							
GBLUP	+Poly[a]	0.59	0.93	0.65	0.91	0.54	0.71	0.15	1.05
	-Poly[b]	0.58	1.05	0.64	1.09	0.54	0.77	0.14	1.08
BR	+Poly[a]	0.64	0.94	0.68	0.93	0.71	0.87	0.15	1.02
	-Poly[b]	0.63	0.98	0.68	1.01	0.69	0.93	0.14	1.04
EM	+Poly[a]	0.64	0.87	0.68	0.92	0.69	0.75	0.15	1.09
	-Poly[b]	0.64	0.98	0.66	1.01	0.67	0.79	0.15	1.09
HB	+Poly[a]	0.64	0.97	0.68	0.90	0.71	0.89	0.15	1.02
	-Poly[b]	0.64	1.06	0.66	0.96	0.69	0.87	0.15	1.02

[a]means polygenic effects are included in the predictions; while [b]means the predictions do not include polygenic effects into the model

Table 6 The accuracy and bias of multi-breeds prediction of GBLUP, BayesR(BR), emBayesR (EM), and HyB_BR (HB)

		Milk yield		Protein yield		Fat%		Fertility	
		Acc.	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
		Holstein and Jersey reference to predict Holstein validation							
GBLUP	+Poly[a]	0.63	0.83	0.65	0.85	0.74	0.85	0.44	1.66
	-Poly[b]	0.62	0.90	0.57	0.88	0.72	0.90	0.42	1.66
BR	+Poly[a]	0.68	0.84	0.68	0.88	0.81	0.90	0.44	1.53
	-Poly[b]	0.67	0.91	0.67	1.03	0.79	0.98	0.42	1.53
EM	+Poly[a]	0.68	0.90	0.68	0.79	0.77	0.83	0.44	1.27
	-Poly[b]	0.65	0.91	0.66	0.85	0.75	0.87	0.44	1.27
HB	+Poly[a]	0.68	0.82	0.67	0.88	0.81	0.94	0.44	1.33
	-Poly[b]	0.67	0.89	0.67	0.95	0.80	1.08	0.44	1.33
		Holstein and Jersey reference to predict Jersey validation							
GBLUP	+Poly[a]	0.64	0.78	0.68	0.85	0.66	0.73	0.24	1.12
	-Poly[b]	0.64	0.90	0.69	1.02	0.64	0.80	0.24	1.12
BR	+Poly[a]	0.69	0.85	0.71	0.99	0.76	0.88	0.26	1.23
	-Poly[b]	0.68	0.95	0.71	1.09	0.74	0.94	0.25	1.24
EM	+Poly[a]	0.66	0.84	0.69	0.71	0.75	0.76	0.23	1.13
	-Poly[b]	0.63	0.86	0.68	0.73	0.70	0.82	0.23	1.13
HB	+Poly[a]	0.71	0.89	0.74	0.94	0.77	0.89	0.26	1.02
	-Poly[b]	0.69	0.98	0.73	1.02	0.73	0.97	0.26	1.02

[a]means polygenic effects are included in the predictions; while [b]means the predictions do not include polygenic effects into the model

Table 7 The accuracy and bias of across-breeds prediction of BayesR, GBLUP, and HyB_BR

	Milk yield		Protein yield		Fat%		Fertility	
	Acc.	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
	Across breeds prediction on Australian red bulls							
GBLUP	0.16	0.54	0.11	0.51	0.32	0.90	0.29	0.97
BayesR	0.22	0.60	0.12	0.49	0.45	0.92	0.27	1.03
EmBayesR	0.24	0.70	0.12	0.42	0.41	0.89	0.29	1.10
HyB_BR	0.23	0.74	0.17	0.49	0.50	0.90	0.30	0.98
	Across breeds prediction on Australian red cows							
GBLUP	0.15	0.71	0.08	0.13	0.50	1.19	0.08	0.79
BayesR	0.26	0.80	0.17	0.51	0.54	0.94	0.08	0.68
EmBayesR	0.24	0.79	0.16	0.53	0.51	0.89	0.08	0.74
HyB_BR	0.26	0.81	0.16	0.57	0.55	0.91	0.08	0.70

BayesR and HyB_BR had a 1 % ~ 6 % superiority of accuracy over GBLUP for Milk yield, Protein yield and Fat%, but had no advantage for fertility. Similarly, for the prediction of Jersey validation with Jersey reference, BayesR and HyB_BR had a consistent advantage over GBLUP for milk production traits, but not for fertility. Especially, for the trait Fat%, BayesR and HyB_BR gave very similar results, with a 17 % increase in accuracy (0.79 vs 0.73 in Holstein and 0.71 vs 0.54 in Jersey) of genomic prediction over GBLUP, as well as a 5 % increase in accuracy over emBayesR. HyB_BR and BayesR also gave regression coefficients closer to one than GBLUP for most traits.

Genomic prediction with a multi-breed reference

When predicting the Holstein or Jersey validation with the combined Holstein and Jersey reference, HyB_BR had the same accuracy as BayesR, Table 5. Compared with GBLUP, BayesR and HyB_BR gave consistently higher accuracy increase for the milk production traits, though this was not observed for fertility. And for the prediction of Jersey validation set, BayesR and HyB_BR improved accuracy for the milk production traits by 11 % compared with GBLUP. The results show that there are small but consistent accuracy improvements as a result of using the multi-breed reference (compare Tables 5 and 6), consistent with the results of Kemper et al. [15] and Hoze et al. [46].

Also, including polygenic effects (estimated using the pedigree) in the model can improve the prediction accuracy by 1 ~ 2 %, at least for milk production traits, Tables 5 and 6. However, for fertility the introduction of polygenic effects for all the prediction methods did not impact the accuracy at all.

Compared with GBLUP and emBayesR, BayesR and HyB_BR gave less biased predictions for milk production traits. However for fertility the regression values far from one indicate bias, from all methods – the low accuracy of fertility phenotypes, including in the validation set, likely contributes to this.

Genomic prediction across breeds

For predicting Australian Red validation bulls (an additional breed to those in the reference set), BayesR and HyB_BR gave higher accuracy than GBLUP for all traits (Table 7).

Across all the prediction results shown in Tables 5, 6 and 7, emBayesR had a 2 % ~ 5 % reduction in accuracy compared with BayesR and HyB_BR for fat%, while BayesR and HyB_BR gave almost identical accuracies in all cases.

Table 8 The number of SNPs in each of four distributions

Traits	The proportion (Pr)	BayesR	emBayesR	HyB_BR
Milk yield	A. 0.01 * σ_g^2	8	6	8
	B. 0.001 * σ_g^2	47	17	327
	C. 0.0001 * σ_g^2	3886	1523	4039
	D. 0	628,062	630,457	627,629
Protein yield	A. 0.01 * σ_g^2	5	4	6
	B. 0.001 * σ_g^2	32	37	297
	C. 0.0001 * σ_g^2	4431	1842	6604
	D. 0	627,535	630,120	625,096
Fat%	A. 0.01 * σ_g^2	23	19	20
	B. 0.001 * σ_g^2	46	206	119
	C. 0.0001 * σ_g^2	2882	1206	1852
	D. 0	629,052	630,572	630,012
Fertility	A. 0.01 * σ_g^2.	10	8	12
	B. 0.001 * σ_g^2	147	114	202
	C. 0.0001 * σ_g^2	3949	8572	7597
	D. 0	627,897	623,309	624,192

Inferred genetic architecture and QTL mapping for dairy production and fertility traits

Bayes R described the genetic architecture of a trait by the posterior proportion of SNPs in each of the four different distributions. Table 8 reported the estimated proportion in each of four distributions from BayesR, emBayesR, and HyB_BR. The number of SNPs falling into the distribution with the largest variance was similar for all three methods. Compared with BayesR, HyB_BR tended to estimate more SNPs (up to 5 %) in the distribution with variance $0.001 * \sigma_g^2$, and $0.0001 * \sigma_g^2$. In contrast to HyB_BR, emBayesR tended to estimate that a higher proportion of SNPs have no effect than does BayesR. This may explain the lower accuracy it achieves.

QTL mapping for dairy production and fertility traits

Both BayesR and HyB_BR estimate the posterior probability that every SNP has a non-zero effect on the trait. This is useful for QTL mapping – SNP with very high posterior probabilities of having a non-zero effect should be strongly associated with causal mutations (e.g. Moser et al. [8], Kemper et al. [15]). Then, QTL mapping from BayesR and HyB_BR can be conducted by plotting the posterior probability of each SNPs having a non-zero

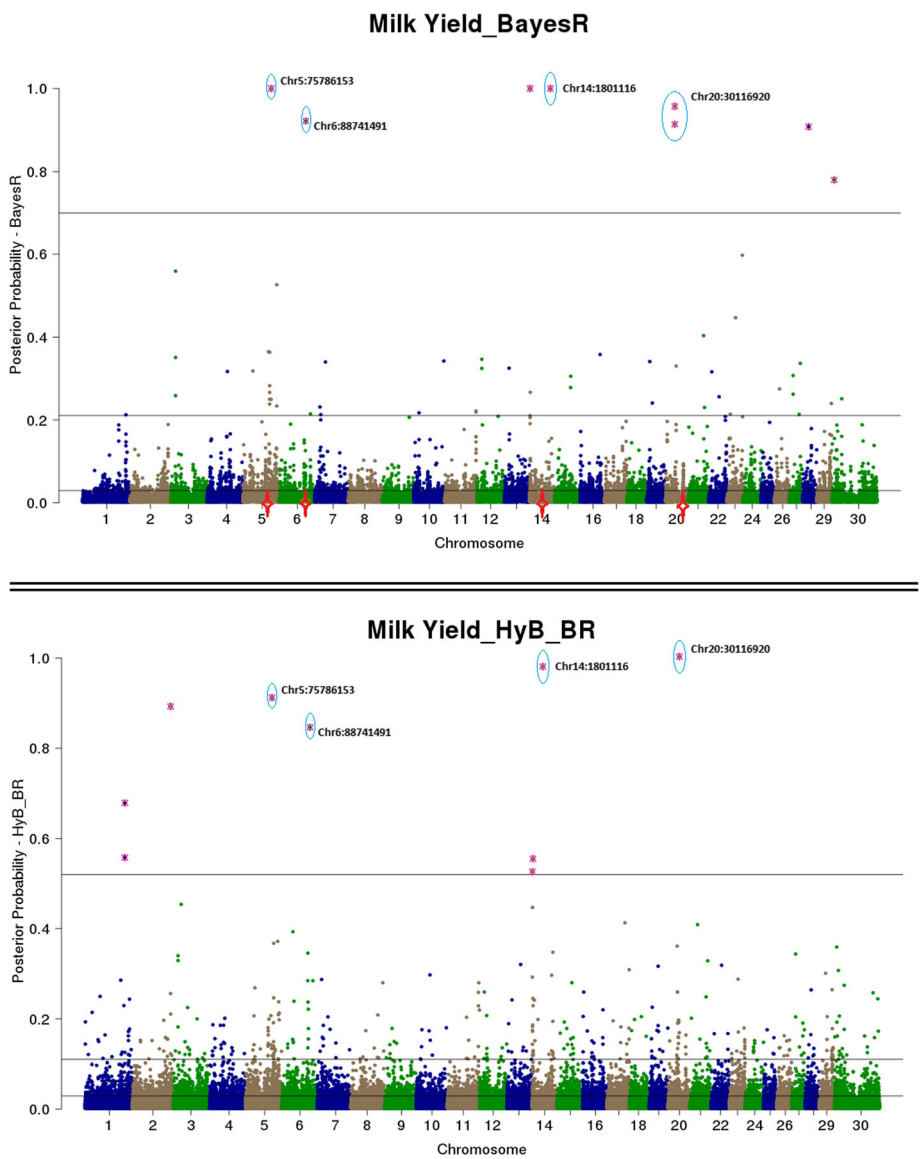

Fig. 4 Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to milk yield. The posterior possibility is calculated based on the sum of the posterior possibilities $P(i, k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i, k)$. The *blue circle* is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes

effect on the trait by genome position, and then comparing the genome location of the effects with a high posterior probability of being in the largest distribution for each method.

The estimated posterior possibilities of all the SNPs (y axis) related to four different traits were plotted according to the positions (base pairs) of SNPs on the whole genome (x axis) in Figs. 4, 5, 6 and 7. The top SNPs with high posterior possibilities were picked up according to the number of SNPs in the variance $0.01 * \sigma_g^2$ (the total number of markers * Pr [4]). Table 9 listed all the top SNPs in the variance related to the previously reported genes with a effect on milk production including CSF2RB [47], GC [48], GHR/CCL28 [18], PAEP [17], MGST1 [49], and DGAT1 [16]. Both BayesR and HyB_BR identified all of these regions, which demonstrated that HyB_BR can perform QTL mapping with similar precision to BayesR. For example, HyB_BR could detect the DGAT1 as well as BayesR shown in Fig. 6 (Fat%).

The application of HyB_BR to predict the risk of Human disease traits and infer genetic architecture for these traits

In the human data, cross validation was used to estimate the accuracy of HyB_BR. As there were 20 replicates of 20/80 split (validation/reference), we evaluated the mean

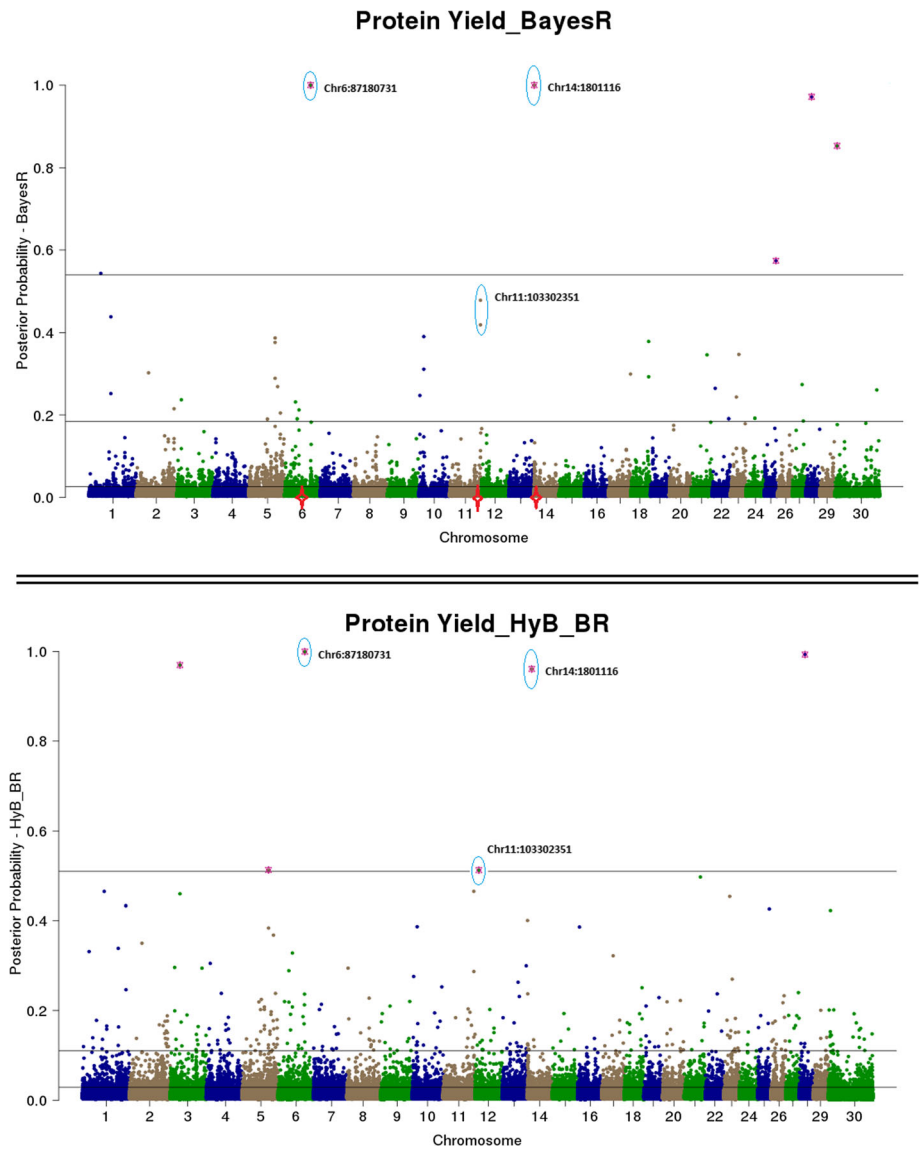

Fig. 5 Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to protein yield. The posterior possibility is calculated based on the sum of the posterior possibilities $P(i, k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i, k)$. The *blue circle* is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes

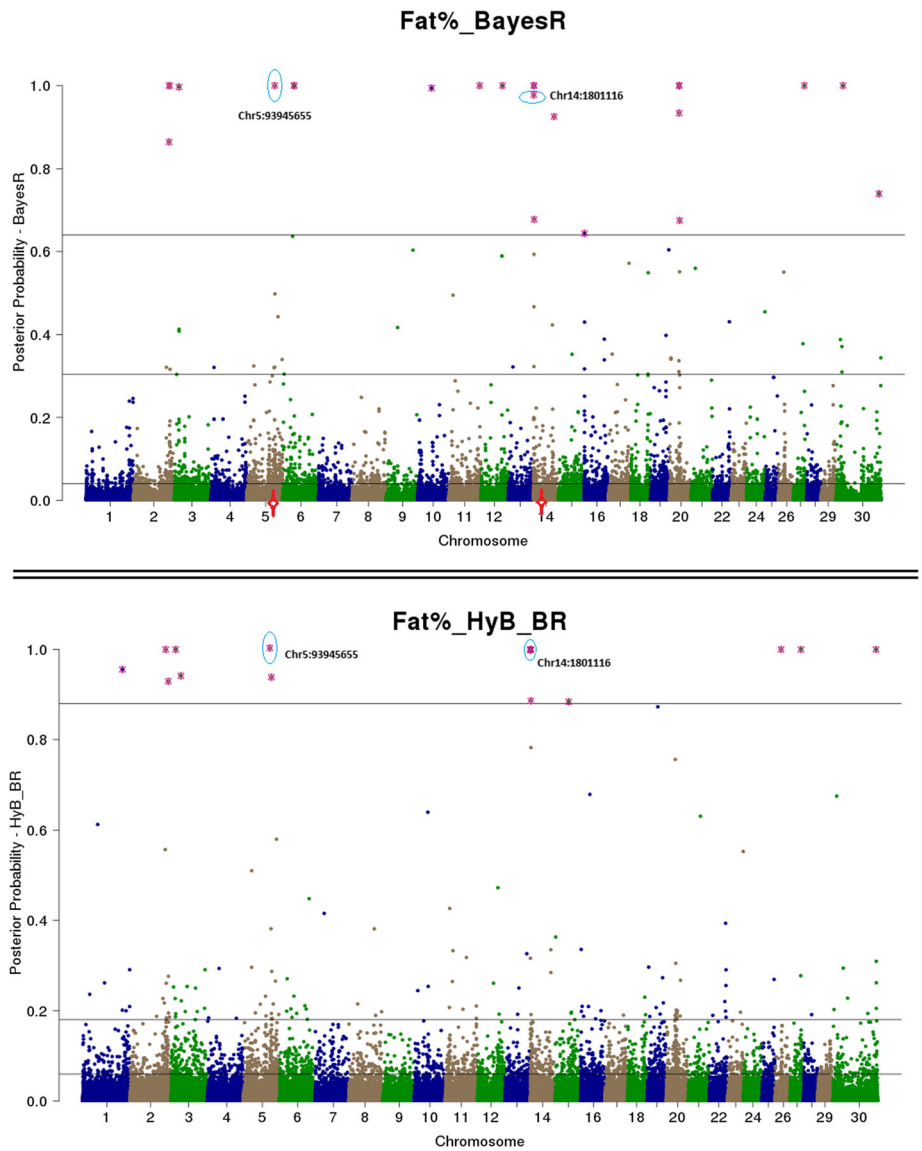

Fig. 6 Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to Fat percent (Fat%). The posterior possibility is calculated based on the sum of the posterior possibilities $P(i, k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i, k)$. The *blue circle* is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes

of the AUC for each disease shown in Table 10. Analysis methods compared were GBLUP implemented in GCTA [44], BayesR from Moser et al. [8], and HyB_BR. The standard deviations of the accuracy (across the 20 replicates) were also listed in the parenthesis of Table 10. HyB_BR and BayesR performed equally well across all seven traits, with the same prediction accuracy for each trait. For the diseases of CD, RA, and T1D, BayesR and HyB_BR had significantly higher accuracy than GBLUP. Especially for T1D, BayesR and HyB_BR could have up to 12 % accuracy increase compared with GBLUP. However, for other traits including BD, CAD, HT, and T2D, BayesR and HyB_BR did not show any superiority over GBLUP. The underlying architecture of these traits might explained this, as suggested by Moser et al. [8]. In detail, for CD, RA and T1D, there are known mutations of moderate to large effect, and the mixture assumptions of BayesR and HyB_BR can take advantage of this. However, for four other diseases including BD, CAD, HT, and T2D, there are no known mutations of moderate to large effect, and this is reflected in the genetic architecture for these diseases inferred by HyB_BR.

The genetic architecture of human disease traits

The inferred genetic architecture was different for each of the seven diseases (Table 11). For example, the

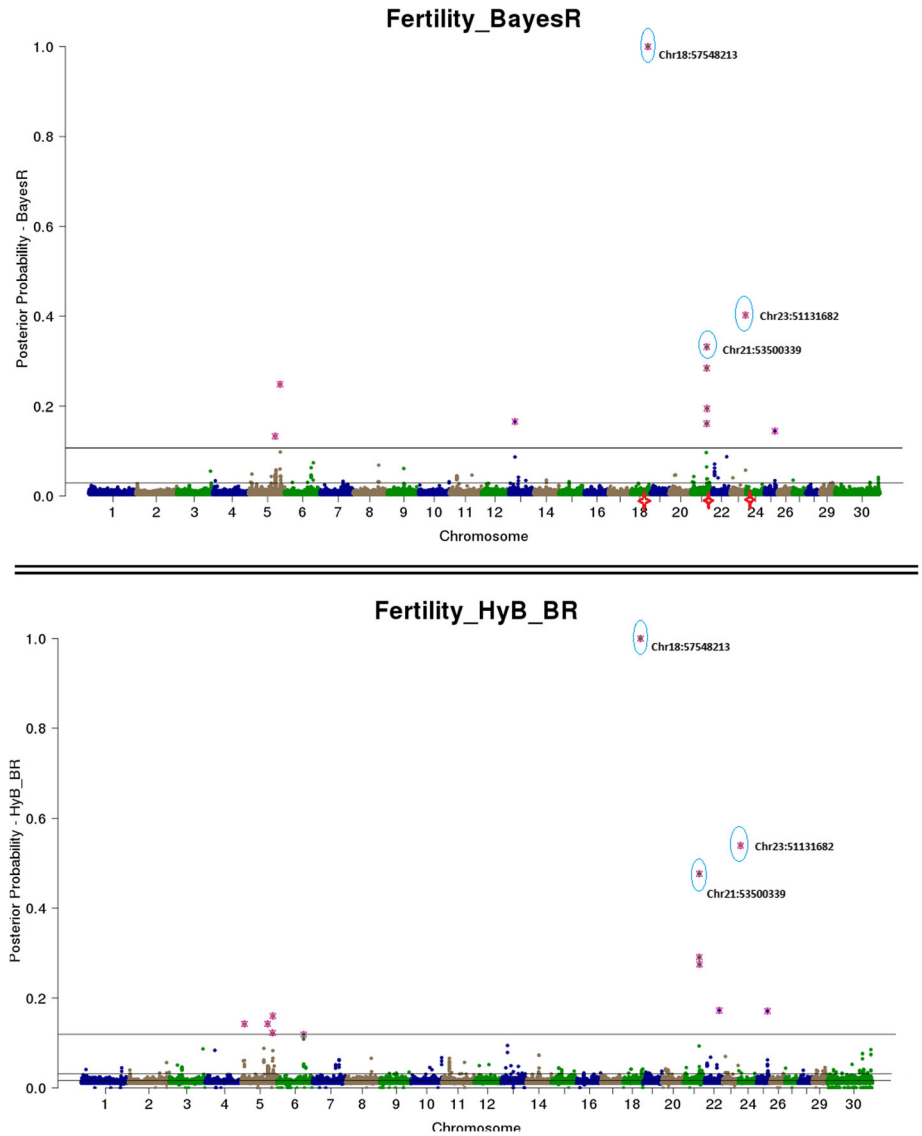

Fig. 7 Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to fertility. The posterior possibility is calculated based on the sum of the posterior possibilities $P(i, k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i, k)$. The *blue circle* is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes

genetic architecture of BD is controlled by many SNPs (9077 for HyB_BR; 9611 for BayesR) with small effects (the variance $0.0001\sigma_g^2$), but just 3 SNPs with large effects (the variance $0.01\sigma_g^2$). These numbers demonstrated the polygenic architecture of BD. On the contrary, for T1D, there was relatively smaller number of SNPs (3544 for HyB_BR; 2750 for BayesR) with small effects but many more SNPs (almost 200) with large effects. The proportion numbers from Fig. 8 also demonstrated this (in accordance with the results from Moser et al. [8]). Large proportion of SNPs with small effects (the variance $0.0001\sigma_g^2$) controlled the polygenic architecture of the diseases BD (98.76 % for HyB_BR; 99.55 % for BayesR), CAD (97.31 % for HyB_BR; 96.8 % for BayesR), HT (96.96 % for HyB_BR; 98.09 % for BayesR), and T2D (95.14 % for HyB_BR; 97.79 % for BayesR). For these diseases, the mixture model of BayesR and HyB_BR did not have much advantage. However, relatively larger proportions of SNPs with moderate effects (the variance $0.001\sigma_g^2$) existed for the traits RA (0.77 % for HyB_BR; 0.93 % for BayesR) and T1D(5.02 % for HyB_BR; 5.54 % for BayesR). For these two traits controlled by major genes, BayesR and HyB_BR gave substantially greater accuracy than GBLUP, which explained the results for prediction accuracy (Table 10).

Table 9 The list of identified causal mutations by both BayesR and HyB_BR

Traits	Loci	Information (known genes)	Range (bp) [Start points ~ End points]
Milk yield	Chr5:75786153	CSF2RB impacting milk yield [47].	[75724620 ~ 75745819]
	Chr6:88741491	GC, encoding the vitamin D binding protein, positively impacting the milk yield [48].	[88695940 ~ 88749180]
	Chr20:30116920	In association with CCL28/GHR impacting milk production [18].	[31890736 ~ 32199996]
Protein yield	Chr6:87180731	CSN1S1 positively impacting protein yield [48].	[87141556 ~ 87159096]
	Chr11:103302351	PAEP impacting protein yield [19].	[103301664 ~ 103306381]
Fat%	Chr5:93945655	MGST1 for Fat percent [49].	[93926791 ~ 3950162]
Fertility	Chr18:57548213	-In association with the gene CEACAM18, Detected by Pryce et al. [50], Cole et al. [51].	~57MBP
	Chr21:53500339	- Control the percentage of unassisted births in first calf heifers [52].	~53MBP
	Chr23:51131682	In the linkage with the known gene GMDS [53].	~51MBP
All the traits	Chr14:1801116	DGAT1 impacting Fat percent [16].	[1795351 ~ 1804562]

Compared with BayesR, HyB_BR detected the same number of SNPs with moderate variance (the variance $0.01 * \sigma_g^2$) but appeared to systematically detect more SNPs in the proportion of small variance (the variance $0.0001 * \sigma_g^2$), similar to the results observed for the comparison of BayesR and HyB_BR in in dairy cattle data (Table 8).

Discussion

We have presented a novel and computationally efficient algorithm termed HyB_BR for simultaneous genomic prediction and QTL mapping. A pure EM algorithm was less accurate for some traits, while pure MCMC requires very long computation times. Therefore, HyB_BR implements the EM algorithm followed by a limited number of MCMC iterations. In this way, the algorithm takes advantage of the features of an EM algorithm (rapid convergence) and the higher accuracy from MCMC implementations in a hybrid scheme. Our accuracies of genomic prediction for complex traits in human and cattle from HyB_BR are almost identical to those from the full MCMC implementation of the Bayesian mixture model, with a 10 fold or greater reduction in computing time required.

For the pure MCMC algorithm, the burn-in stage can account for up to 50 % of the total running time. One of the key advantages of HyB_BR is that the EM module effectively replaces the burn-in cycles that are usually required for MCMC. Based on the starting point from EM (with very limited number of iterations; less than 500 iterations), the running time of HyB_BR can be much reduced.

The pure EM algorithm, EmBayesR [30] has been demonstrated to be much faster than BayesR, but had lower accuracy for some traits, particularly those with mutations of moderate to large effect. For example, when implemented on the trait fat% in dairy cattle, emBayesR had a decreased accuracy of 5 % ~ 7 % compared to BayesR. One possible explanation is that emBayesR shrinks SNP effects too much (shown in Table 8). This could be because the PEV that is used to account for the error of the effects of all the other SNPs while estimating the effect of the current SNP is only an approximation. The introduction of PEV correction is based on one observation: previous fast algorithm stud-

Table 10 The prediction performance evaluated by the Area under curve (AUC) of GBLUP, BayesR and HyB_BR on seven diseases

Diseases	GBLUP		BayesR		HyB_BR	
	AUC	h^2	AUC	h^2	AUC	h^2
BD	0.63(0.0135)	0.71	0.63(0.0131)	0.63	0.64(0.0174)	0.63
CAD	0.58(0.0116)	0.38	0.59(0.0118)	0.38	0.58(0.0131)	0.38
CD	0.60(0.0134)	0.69	0.65(0.0159)	0.61	0.65(0.0158)	0.61
HT	0.58(0.0125)	0.53	0.58(0.0131)	0.52	0.58(0.0140)	0.51
RA	0.58(0.0109)	0.50	0.70(0.0104)	0.45	0.70(0.0107)	0.45
T1D	0.64(0.0133)	0.66	0.86(0.0099)	0.63	0.86(0.0102)	0.63
T2D	0.59(0.0139)	0.59	0.60(0.0117)	0.52	0.60(0.0122)	0.52

the heritability (h^2) is estimated by the equation $h^2 = \frac{\sigma_g^2}{(\sigma_e^2+\sigma_g^2)}$; σ_e^2 is derived separately by three methods; fixed genetic variance of σ_g^2 for BayesR and HyB_BR is obtained from GCTA

Table 11 The number of SNPs in each proportion of four distributions estimated by BayesR, and HyB_BR on seven human diseases

Diseases	BayesR				HyB_BR			
	Pr[54]	Pr[2]	Pr[3]	Pr[4]	Pr[54]	Pr[2]	Pr[3]	Pr[4]
BD	282,843	9611	39	3	283,306	9077	110	3
CAD	289,491	6892	214	13	289,203	7211	183	13
CD	294,423	6878	269	9	294,463	6576	331	9
HT	286,152	8094	150	8	286,160	7993	243	8
RA	291,401	4172	275	42	290,420	5025	403	42
T1D	293,366	2607	54	200	292,523	3396	104	207
T2D	286,489	7972	173	7	288,365	5971	298	7

ies (especially Iterative conditional expectation algorithms) assumed the effects of the other SNPs were estimated perfectly while estimating the effect of the current SNP, leading to poor performance [30]. Therefore, EmBayesR and the EM part of HyB_BR allow for the errors in the effect of other SNPs and other location parameters by using the PEV. The calculation of the PEV from GBLUP is carried out before the iterations to estimate the effects of each SNP. And since the normal priors from GBLUP model do not allow for SNPs of moderate to large effects, such PEV calculation is an approximation and this may be one reason for loss of accuracy in the EM. To deal with this, HyB_BR further implements a small number of MCMC iterations to improve the outcome of pure EM steps.

HyB_BR has three advantages. First, as the size of genomic data increases, the computational efficiency of HyB_BR without burn-in stage (a small number of $O(mn)$ iterations), is greater than BayesR by full MCMC. And when implemented with the speed-up scheme described in the methods, computational time can be reduced even further, by sampling a reduced set of SNPs

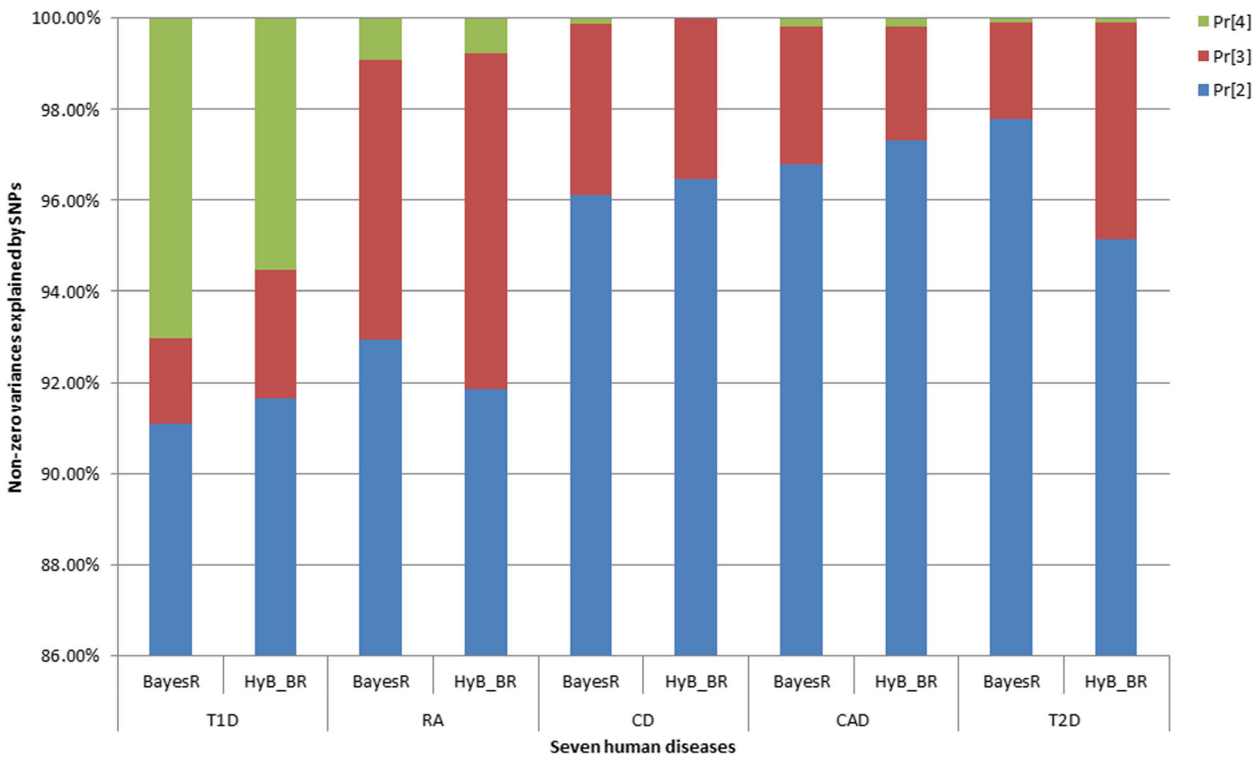

Fig. 8 The inferred genetic architecture of seven human diseases from BayesR and HyB_BR. The *blue bar* is the proportion of SNPs in Pr [2] (with the variance 0.0001 * σ_g^2), which is estimated by the number of SNP in Pr [2] divided by the total number of SNPs with nonzero variance. The *red bar* is the proportion of SNPs with the variance 0.001 * σ_g^2, estimated by the number of SNP in Pr [3] divided by the total number of SNPs with nonzero variance. The *green bar* is the proportion of SNPs with the variances 0.01 * σ_g^2, estimated by the number of SNPs in Pr [2] divided by the total number of SNPs with nonzero variance

in the MCMC module, apparently with no loss of accuracy (but critically the information from the SNPs that are not sampled remains in the posterior proportions of SNPs in each distribution). Second, the prediction accuracy of HyB_BR is comparable to BayesR in all cases including dairy cattle and human disease prediction shown in Tables 5, 6 and 7, and Table 10. Third, HyB_BR, like BayesR, is flexible with respect to the genetic architecture of complex traits. As shown in Tables 5, 6 and 7, HyB_BR performs well on four different complex traits, with architecture ranging from highly polygenic architecture to genetic architecture controlled by major genes. In addition to the prediction on the continuous quantitative traits of dairy cattle, the investigation on the risk prediction of seven case/control human diseases with binary 0/1 phenotypes shows HyB_BR and BayesR perform on this type of data, Table 10. Finally, the posterior probabilities of SNP having a nonzero effect from HyB_BR can be used for QTL mapping, Fig. 6.

Implementing genomic prediction methods with whole genome sequence data may improve the prediction accuracy and accelerate the discovery of causal variants. However, for this to occur, more computationally efficient genomic prediction algorithms are required. Compared with BayesR, the predicted time of HyB_BR on different number of markers with the same reference phenotypesis listed in Table 12. The time is estimated linearly on the number of markers and individuals. When the number of markers reaches 30 million (the number of variants discovered in the 1000 bull genomes project, Daetwyler et al. [28]), the running time of BayesR is around 34,170 h, which is impractical. On the contrary, on the same data with 30 million of variants, HyB_BR is predicted to require 2010 h. It may be possible to reduce this further by optimising the code even more. Therefore, as the size of genomic data increases, HyB_BR will remain feasible well beyond the point where the use of BayesR is impractical.

While HyB_BR performs well with computational efficiency and robust prediction accuracy, there are at least still two strategies that could be used to further improve efficiency. There is one key part of EM module that consumes running time and memory: the calculation of $tr(\mathbf{E}^{-1}\mathbf{Z_i}\mathbf{Z_i'}\mathbf{E}^{-1}\mathbf{PEV})$ for each SNP in front of EM iterations. In detail, the calculation of $tr(\mathbf{E}^{-1}\mathbf{Z_i}\mathbf{Z_i'}\mathbf{E}^{-1}\mathbf{PEV})$ requires the time complexity of $O\left(\frac{1}{2}mn^2\right)$, which accounts for almost 2/3 of the computational time required for the EM module even though the calculation is made in front of the EM iterations. Therefore, a future task is to implement a multi-threaded version to improve speed. The threshold of limiting the number of SNPs to be updated requires further study. Currently we define the threshold as **T**: if$(P(i, 1) > 0.9)$, which is applicable for the current data. However, it's uncertain whether or not such a threshold is suitable for other types of data.

Table 12 The predicted computational time (in hours) of HyB_BR and BayesR on high density data with different number of variants and the same number of individuals (16,214)

	Different number of markers			
	800 K SNP panel	1 million	2 million	30 millions
BayesR	720 h	1139 h	2278 h	34,170 h
HyB_BR	42 h	67 h	134 h	2010 h

HyB_BR has some features in common with other mixture methods such as BSLMM [6], and BOLT-LMM [25]. All of these methods declared the merit of computational efficiency with time complexity $O(mn)$ but under different mixture models. In detail, BSLMM assumed a large proportion of SNPs with small effects (under BLUP models), while others had large effects (under Bayesian sparse regression models; the mixture of two normal priors). Due to limited number of SNPs implemented for MCMC sampling (large proportion of SNPs are under GBLUP models), BSLMM could be computationally efficient. However, compared with the mixture of four normal distributions by BayesR which provided great flexibility with respect genetic architecture, the flexibility of BSLMM with respect to different genetic architectures required further investigation. Another algorithm is BOLT-LMM, which has been developed mainly for the association studies. BOLT-LMM incorporated Bayesian mixture models to improve the power of GWAS with appealing outcomes. Instead of MCMC sampling, BOLT-LMM implemented iterative conditional expectation (ICE) algorithm on a mixture of two normal distributions to improve the computational speed with the approximated computational complexity $O(mn)$. There could be three limitations with this method: 1) ICE algorithms did not account for the PEVs from all other SNP effects during the estimation of current SNP effect. On practical data sets, ICE could lead to the loss of prediction accuracy. BOLT-LMM introduced LD score regression technique to calibrate the prediction errors. However, since the calibrating factor was constant across all the SNPs (the prediction error variance regarding each SNP differs according to our equation $(\mathbf{E}^{-1}\mathbf{Z_i}\mathbf{Z_i'}\mathbf{E}^{-1}\mathbf{PEV})$), such calibration scheme seem not to be effective to solve the problem. 2) The leave-one-chromosome-out scheme implemented in BOLT-LMM might perform well for GWAS but not be suitable for simultaneous genomic prediction. 3) BOLT-LMM treated each SNP effect as a fixed effect for the association statistics. This combined with the stringent significance threshold for multiple testing, leaded to the over-estimation for SNP effects. Another efficient method for genomic prediction termed MultiBLUP [7] introduced SNPs clusters into BLUP models according

to its adaptive algorithm. For each SNP class, the linear combination models (using genomic relationship matrix) similar to GBLUP were implemented. MultiBLUP has been demonstrated to be computationally efficient with robust prediction accuracy in the human data sets. However, when moved to dairy cattle genomic data sets, there is long Linkage disequilibrium (LD) between markers, which might be easily broken up by multiBLUP models.

Conclusions

In summary, HyB_BR is a computationally efficient method for simultaneous genomic prediction, QTL mapping and inference of genetic architecture. The hybrid scheme of MCMC and EM decreases computational time by a factor of at least 10 fold with no reduction in prediction accuracy. The HyB_BR algorithm makes simultaneous genomic prediction, QTL mapping and inference of genetic architecture feasible in extremely large genomic data sets including whole genome sequence data.

Additional files

Additional file 1: PEV calculation from GBLUP. (DOCX 16 kb)
Additional file 2: Calculation of $P(i, k)$. (DOCX 21 kb)

Acknowledgements
The authors acknowledge the support from Dairy Futures CRC project.

Availability of data and materials
The WTCCC data are available to researchers by application to the Welcome Trust Case Control Consortium Data Access Committee (http://www.wtccc.org.uk/info/access_to_data_samples.html), or contact egaadmin@ebi.ac.uk. Application is required to ensure proper protection of confidentiality of the participants.
For dairy cattle data, we can provide meta-analysis data related to our paper which can be easily used to conduct the analysis by other researchers.
The HyB_BR compiled program is available for request for non-commercial research.

Authors' contributions
BJH and Y-PPC supervised this project; TW developed HyB_BR algorithm, analysed the data and drafted the manuscript; BJH gave important instructions on organizing and revising the manuscript. MEG contributed the valuable idea about hybrid scheme; PJB provided help with C++ programming. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC, Australia. [2]Biosciences Research, Department of Economic Development, Jobs, Transport and Resources, Bundoora, Melbourne, VIC, Australia. [3]Dairy Futures Cooperative Research Centre, Melbourne, VIC, Australia. [4]School of Applied Systems Biology, La Trobe University, Melbourne, VIC, Australia. [5]Faculty of Veterinary and Agricultural Science, University of Melbourne, Melbourne, VIC, Australia. [6]Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, University of Queensland, Queensland, Australia.

Received: 23 March 2016 Accepted: 10 September 2016
Published online: 21 September 2016

References
1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
2. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10(6):381–91.
3. Meuwissen T, Hayes B, Goddard M. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013;1(1):221–37.
4. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–9.
5. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010;11(12):880–6.
6. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.
7. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24(9):1550–7.
8. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11(4):e1004969.
9. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
10. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, Sonstegard TS, Connor EE, Winters M, van Kaam JBCHM, et al. Genomic imputation and evaluation using high-density holstein genotypes. J Dairy Sci. 2013;96(1):668–78.
11. Wolc A, Zhao HH, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Stricker C, Habier D, Fernando RL, et al. Response and inbreeding from a genomic selection experiment in layer chickens. Genet Sel Evol. 2015;47:59.
12. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 2013;194(3):573–96.
13. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinf. 2011;12(1):1–12.
14. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114–29.
15. Kemper KE, Reich CM, Bowman PJ, vander Jagt CJ, Chamberlain AJ, Mason BA, Hayes BJ, Goddard ME. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 2015;47:29.
16. Grisart B, Coppieters W, Farnir F, Karim LCF, Berzi P, Cambisano N, Mni M, Reid S, Simon P, et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 2002;12(2):222–31.
17. Ng-Kwai-Hang K. A review of the relationship between milk protein polymorphism and milk composition/milk production. In: Proceedings of the international dairy federation seminar: 25–27 febuary, 1997 1997; palmerston north, New Zealand. 1997. p. 22–37.
18. Blott S, Kim J-J, Moisio S, Schmidt-Küntzel A, Cornet A, Berzi P, Cambisano N, Ford C, Grisart B, Johnson D, et al. Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics. 2003;163(1):253–66.
19. Wang X, Wurmser C, Pausch H, Jung S, Reinhardt F, Tetens J, Thaller G, Fries R. Identification and dissection of four major QTL affecting milk Fat content in the German holstein-friesian population. PLoS One. 2012;7(7):e40711.
20. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42(4):355–60.
21. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Meth. 2011;8(10):833–5.
22. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4.

23. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. Nat Meth. 2012;9(6):525–6.
24. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46(2):100–6.
25. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47(3):284–90.
26. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46(11):1173–86.
27. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
28. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46(8):858–65.
29. Calus MPL. Right-hand-side updating for fast computing of genomic breeding values. Genet Sel Evol. 2014;46:24.
30. Wang T, Chen Y-PP, Goddard ME, Meuwissen THE, Kemper KE, Hayes BJ. A computationally efficient algorithm for genomic prediction using a Bayesian model. Genet Sel Evol. 2015;47:34.
31. Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol. 2009;41:2.
32. Yu X, Meuwissen THE. Using the pareto principle in genome-wide breeding value estimation. Genet Sel Evol. 2011;43:35.
33. Shepherd RK, Meuwissen THE, Woolliams JA. Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. BMC Bioinf. 2010;11(1):1–12.
34. Hayashi T, Iwata H. EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet. 2010;11(1):1–9.
35. Sun X, Qu L, Garrick DJ, Dekkers JCM, Fernando RL. A fast EM algorithm for BayesA-like prediction of genomic breeding values. PLoS One. 2012;7(11):e49157.
36. Garrick D, Taylor J, Fernando R. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009;41(1):55.
37. Henderson C. Application of linear models in animal breeding. Canada: University of Guelph; 1984.
38. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210–23.
39. Haile-Mariam M, Bowman PJ, Pryce JE. Genetic analyses of fertility and predictor traits in Holstein herds with low and high mean calving intervals and in Jersey herds. J Dairy Sci. 2013;96(1):655–67.
40. Haile-Mariam M, Pryce JE, Schrooten C, Hayes BJ. Including overseas performance information in genomic evaluations of Australian dairy cattle. J Dairy Sci. 2015;98(5):3443–59.
41. Gilmour A, Cullis B, Welham S, Thompson R. ASReml reference manual 2nd edition, NSW agriculture biometrical bulletin 3. 2002.
42. Lee Sang H, Wray Naomi R, Goddard Michael E, Visscher Peter M. Estimating missing heritability for disease from genome-wide association studies. Am J Human Gen. 2011;88(3):294–305.
43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: a tool Set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
44. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
45. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010;6(2):e1000864.
46. Hozé C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J Dairy Sci. 2014;97(6):3918–29.
47. Chamberlain AJ, Vander Jagt CJ, Hayes BJ, Khansefid M, Marett LC, Millen CA, Nguyen TTT, Goddard ME. Extensive variation between tissues in allele specific expression in an outbred mammal. BMC Genomics. 2015;16(1):1–20.
48. Sanders K, Bennewitz J, Reinsch N, Thaller G, Prinzenberg EM, Kühn C, Kalm E. Characterization of the DGAT1 mutations and the CSN1S1 promoter in the German angeln dairy cattle population. J Dairy Sci. 2006;89(8):3164–74.
49. Raven L-A, Cocks BG, Kemper KE, Chamberlain AJ, Jagt CJ, Goddard ME, Hayes BJ. Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. Mamm Genome. 2015;27(1):81–97.
50. Pryce JE, Bolormaa S, Chamberlain AJ, Bowman PJ, Savin K, Goddard ME, Hayes BJ. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J Dairy Sci. 2010;93(7):3331–45.
51. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK, et al. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics. 2011;12(1):1–17.
52. McClure MC, Morsci NS, Schnabel RD, Kim JW, Yao P, Rolf MM, McKay SD, Gregg SJ, Chapple RH, Northcutt SL, et al. A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. Anim Genet. 2010;41(6):597–607.
53. Wickramasinghe S, Hua S, Rincon G, Islas-Trejo A, German JB, Lebrilla CB, Medrano JF. Transcriptome profiling of bovine milk oligosaccharide metabolism genes using RNA-sequencing. PLoS One. 2011;6(4):e18895.
54. Consortium TGP. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central