BMC Genetics

METHODOLOGY ARTICLE Open Access

Multi-population genomic prediction using a multi-task Bayesian learning model

Liuhong Chen[1*], Changxi Li[1,3], Stephen Miller[2] and Flavio Schenkel[2]

Abstract

Background: Genomic prediction in multiple populations can be viewed as a multi-task learning problem where tasks are to derive prediction equations for each population and multi-task learning property can be improved by sharing information across populations. The goal of this study was to develop a multi-task Bayesian learning model for multi-population genomic prediction with a strategy to effectively share information across populations. Simulation studies and real data from Holstein and Ayrshire dairy breeds with phenotypes on five milk production traits were used to evaluate the proposed multi-task Bayesian learning model and compare with a single-task model and a simple data pooling method.

Results: A multi-task Bayesian learning model was proposed for multi-population genomic prediction. Information was shared across populations through a common set of latent indicator variables while SNP effects were allowed to vary in different populations. Both simulation studies and real data analysis showed the effectiveness of the multi-task model in improving genomic prediction accuracy for the smaller Ayshire breed. Simulation studies suggested that the multi-task model was most effective when the number of QTL was small ($n = 20$), with an increase of accuracy by up to 0.09 when QTL effects were lowly correlated between two populations ($\rho = 0.2$), and up to 0.16 when QTL effects were highly correlated ($\rho = 0.8$). When QTL genotypes were included for training and validation, the improvements were 0.16 and 0.22, respectively, for scenarios of the low and high correlation of QTL effects between two populations. When the number of QTL was large ($n = 200$), improvement was small with a maximum of 0.02 when QTL genotypes were not included for genomic prediction. Reduction in accuracy was observed for the simple pooling method when the number of QTL was small and correlation of QTL effects between the two populations was low. For the real data, the multi-task model achieved an increase of accuracy between 0 and 0.07 in the Ayrshire validation set when 28,206 SNPs were used, while the simple data pooling method resulted in a reduction of accuracy for all traits except for protein percentage. When 246,668 SNPs were used, the accuracy achieved from the multi-task model increased by 0 to 0.03, while using the pooling method resulted in a reduction of accuracy by 0.01 to 0.09. In the Holstein population, the three methods had similar performance.

Conclusions: Results in this study suggest that the proposed multi-task Bayesian learning model for multi-population genomic prediction is effective and has the potential to improve the accuracy of genomic prediction.

Keywords: Multi-task learning, Bayesian model, Multi-population, Genomic prediction, Stochastic search variable selection

Background

Genomic prediction has become a new tool for selection of candidates based on genomic estimated breeding values (GEBV) through the use of dense markers covering the whole genome [1]. To predict GEBV, a training data set with genotypes and phenotypes is used to derive the prediction equations, where all marker effects are estimated simultaneously. GEBV for selection candidates that have genotypes are then predicted by summing up all the marker effects. The accuracy of GEBV is affected by several factors [2,3], of which the number of individuals in the training data set and the marker density are of crucial importance [2,3].

In Holstein dairy cattle, genomic prediction has been successfully applied using the Illumina BovineSNP50 single nucleotide polymorphism (SNP) panel [4,5]. For

* Correspondence: liuhong@ualberta.ca
[1]Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G 2P5, Canada
Full list of author information is available at the end of the article

smaller populations such as Ayrshire in dairy cattle, acquisition of a large number of animals to be included in the training data set for genomic prediction still remains a challenge. One strategy is to combine data of the small populations with data from other populations to increase the size of the training set. However, simply pooling data from different populations may result in unfavorable accuracies if the marker density is low or the populations have diverged for a long time [6-8]. Increasing the marker density is one of the possible solutions because the linkage disequilibrium (LD) phase persistence between markers and quantitative trait loci (QTL) among different populations would likely to improve. However, a recent study in Jersey and Holstein dairy cattle reported a very limited advantage by using a very high density SNP panel [9]. Few studies have been dedicated to research new methods or strategies other than simply pooling data together for genomic prediction. Brondum et al. [10] proposed an approach called BayesRS for multi-population genomic prediction, where a location specific genetic variance derived in one population were used as priors for another population. They found that for some traits, BayesRS might be advantageous compared to the approach of simply pooling training data sets for distantly-related populations; but for closely related populations the method did not perform better than simply pooling data together.

Multi-task learning, the term first coined by R Caruana [11], aims to improve learning performance by simultaneously learning from related tasks. In text and speech recognition, image reconstruction and many other areas where data are collected from multiple sources, multi-task learning has been successfully applied [12-14]. Recently, the multi-task learning has attracted a growing interest in biological science for sequence and gene expression analyses [15-17] as well as genome-wide association studies (GWAS) [18]. To our knowledge, the application of multi-task learning in genomic selection has not been reported so far.

Bayesian learning models using stochastic search variable selection (SSVS) has been widely used and proven effective for genomic prediction in a single population [1,19-21]. Normally, SSVS uses some types of spike and slab distributions as priors for SNP effects. A latent indicator variable (0 or 1) is associated with each SNP, with 0 indicating that the SNP is irrelevant to the trait and is excluded from the model, and 1 indicating that the SNP is associated to the trait phenotype and is included in the model. In this study, it was assumed that multiple populations share the same set of latent indicator variables which can be learned jointly. The goal was to develop a multi-task Bayesian learning model for multi-population genomic prediction and to evaluate its performance on both simulated and real data.

Methods

In this section, single-task Bayesian learning model, the simple data pooling method, and the multi-task Bayesian learning model were introduced. A Gibbs sampling algorithm was designed to implement the multi-task Bayesian learning model. Monte Carlo simulation studies were conducted to evaluate the performance of the proposed multi-task model. Real data on five milk production traits from Holstein and Ayrshire dairy breeds were also used to test the effectiveness of the multi-task model.

Single-task bayesian learning model

In a single reference population of n animals with genotypes on m SNP markers, the statistical model can be written as:

$$y_i = \mu + \sum_{j=1}^{m} x_{ij} a_j + e_i,$$

Where y_i is the phenotypic value for the i^{th} animal (i = 1,...,n), μ is the intercept, x_{ij} is the genotype for the j^{th} SNP locus (j = 1,...,m) of the i^{th} animal, which is coded 0, 1 or 2, depending on the number of copies from a specified allele, a_j is the regression coefficient for the j^{th} SNP (allele substitution effect), and e_i is the random residual error.

A flat prior distribution is assigned to μ. a_j is assumed a mixture of a normal distribution $N\left(0, \sigma_a^2\right)$ and a point mass density at zero (denoted by a Dirac delta function $\delta_0(a_j)$ hereinafter). The weights for the two distributions are (1-w) and w, respectively, so that $\left(a_j | w, \sigma_a^2\right)\tilde{}(1-w)$ $N\left(0, \sigma_a^2\right) + w\delta_0\left(a_j\right)$. w follows a uniform prior distribution. A latent indicator variable γ_i is introduced for each SNP, so that when γ_i=*1* , $a_j\tilde{}N\left(0, \sigma_a^2\right)$, and when γ_i=*0* , a_j=0. Prior distribution for each γ_i is assumed i.i.d. and follows Bernoulli distribution with probability (1-w). So the joint prior density for γ is $f(\gamma | w) = \prod_j w^{(1-\gamma_j)}(1-w)^{\gamma_j}$.

Residual errors are assumed from a multivariate normal distribution $N\left(0, I\sigma_e^2\right)$. The prior distribution for σ_a^2 $\left(\sigma_e^2\right)$ is a scaled inverse Chi-square distribution with degree of freedom $\nu_a(\nu_e)$ and scale factor $s_a^2\left(s_e^2\right)$.

Simple data pooling method

Suppose animals are from a number of c different populations. In a simple data pooling method, animals from multiple populations are pooled together to form a single training data set. It is assumed that the population origin for each individual is known prior to the analysis. Population origin is included as a fixed effect. The effect of each SNP is assumed to be the same across populations.

Multi-task Bayesian learning model

For c populations with n_k animals in the k-th population, the statistical model can be written as:

$$y_{ik} = \mu_k + \sum_{j=1}^{m} x_{ijk} a_{jk} + e_{ik} \quad (i = 1, \cdots, n_k \text{ and } k = 1, \cdots, c),$$

In matrix notation, this can be written as:

$$\mathbf{y_k} = \boldsymbol{\mu}_\mathbf{k}\mathbf{1} + \mathbf{X_k a_k} + \mathbf{e_k},$$

where y_{ik} is the phenotypic value for the i^{th} animal in the k^{th} population; μ_k is the general mean of population k; x_{ijk} is the genotype for the j^{th} SNP locus of the i^{th} animal in the k^{th} population; a_{jk} is the j^{th} SNP effect in population k, and e_{ik} is the random residual effect; and in the matrix notation, $\mathrm{y_k}$, $\mathrm{a_k}$, and $\mathrm{e_k}$ are vectors of phenotypic values, SNP effects, and residual effects, respectively; 1 is a vector with all elements set to 1; and $\mathrm{X_k}$ is the design matrix relating $\mathrm{y_k}$ to $\mathrm{a_k}$. In the model, a_{jk} allows the j^{th} SNP effect to have a different value in population k. To share information among different populations, a common latent indicator variable indicating whether SNP j is associated with a QTL is used across populations. Accommodating these features into a Bayesian model produces the multi-task Bayesian learning model.

The following prior distributions for the unknown parameters and hyper-parameters are assumed in the multi-task Bayesian learning model:

$$\mu_k \sim \text{flat distribution},$$

$$\left(a_{jk}|\gamma_j, \sigma^2_{ak}\right) \sim \gamma_j N\left(0, \sigma^2_{ak}\right) + \left(1-\gamma_j\right)\delta_0\left(a_{jk}\right),$$

$$\left(\gamma_j|w\right) \sim w^{1-\gamma_j}(1-w)^{\gamma_j},$$

$$w \sim \mathit{Uniform}\,(0,1),$$

$$\left(\sigma^2_{ak}|\nu_{ak}, s^2_{ak}\right) \sim \frac{\left(\nu_{ak}s^2_{ak}/2\right)^{\nu_{ak}/2}}{\Gamma(\nu_{ak}/2)}\left(\sigma^2_{ak}\right)^{-(1+\nu_{ak})/2}\exp\left(-\frac{\nu_{ak}s^2_{ak}}{2\sigma^2_{ak}}\right),$$

$$\left(\mathbf{e_k}|\sigma^2_{ek}\right) \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_{ek}\right),$$

$$\left(\sigma^2_{ek}|\nu_{ek}, s^2_{ek}\right) \sim \frac{\left(\nu_{ek}s^2_{ek}/2\right)^{\nu_{ek}/2}}{\Gamma(\nu_{ek}/2)}\left(\sigma^2_{ek}\right)^{-(1+\nu_{ek})/2}\exp\left(-\frac{\nu_{ek}s^2_{ek}}{2\sigma^2_{ek}}\right).$$

The likelihood function of the whole data given all the parameters in the model is:

$$\prod_{k=1}^{c}\left(2\pi\sigma^2_{ek}\right)^{-\frac{n_k}{2}}\exp\left[-\frac{1}{2}\sigma^{-2}_{ek}\left(\mathbf{y}_k-\boldsymbol{\mu}_k-\mathbf{X}_k\mathbf{a}_k\right)'\left(\mathbf{y}_k-\boldsymbol{\mu}_k-\mathbf{X}_k\mathbf{a}_k\right)\right]$$

So the joint posterior density function is:

$$f\left(\sigma^2_a, \sigma^2_e, w, \gamma, a, \mu|y\right) \propto \prod_{k=1}^{c}\left\{\left(\sigma^2_{ek}\right)^{-\frac{\nu_{ek}+n_k}{2}-1}\exp\left[-\frac{\left(\mathbf{y}_k-\boldsymbol{\mu}_k-\mathbf{X}_k\mathbf{a}_k\right)'\left(\mathbf{y}_k-\boldsymbol{\mu}_k-\mathbf{X}_k\mathbf{a}_k\right)+\nu_{ek}s^2_{ek}}{2\sigma^2_{ek}}\right]\left(\sigma^2_{ak}\right)^{-\frac{\nu_{ak}}{2}-1}\exp\left(-\frac{\nu_{ak}s^2_{ak}}{2\sigma^2_{ak}}\right)\right\}\ \prod_{k=1}^{c}\prod_{j=1}^{m}\left\{\left[\gamma_j\left(\sigma^2_{ak}\right)^{-\frac{1}{2}}\exp\left(-\frac{a^2_{jk}}{2\sigma^2_{ak}}\right)+\left(1-\gamma_j\right)\delta_0\left(a_{jk}\right)\right]w^{\left(1-\gamma_j\right)}(1-w)^{\gamma_j}\right\}$$

Gibbs sampling algorithm

A Gibbs sampling algorithm was designed to draw samples for unknown (hyper-) parameters from their full conditional posterior distributions. To avoid reducibility of Markov chain, γ_j and a_{jk} are jointly sampled by first sampling γ_j from $f\left(\gamma_j|\boldsymbol{\theta}_{j^-}, \boldsymbol{y}\right)$ followed by sampling a_{jk} from $f\left(a_{jk}|\gamma_j, \theta_{j^-}, y\right)$, where θ_{j^-} represents all parameters except γ_j and a_{jk}. Full conditional posterior distributions for μ_k, w, σ^2_{ak} and σ^2_{ek} can be derived by picking up the relevant parts from the joint posterior distribution. Derivation for the density function $f\left(\gamma_j|\boldsymbol{\theta}_{j^-}, \boldsymbol{y}\right)$ and sampling of γ_j are described in the Appendix. The Gibbs sampling steps are described as below:

Step 1. Initialize the parameters $w, \gamma, \sigma^2_{ak}, \sigma^2_{ek}, \mu_k$ and a_k.
Step 2. For $j=1,\cdots, m$

a. Sample γ_j from Bernoulli distribution with probability $1/(1+q_j)$,

$$q_j = \frac{w}{1-w}\sqrt{\prod_k\left(\mathbf{x}'_{jk}\mathbf{x}_{jk}\frac{\sigma^2_{ak}}{\sigma^2_{ek}}+1\right)}\exp\left(-\frac{1}{2}\sum_k\frac{\hat{\mu}^2_{a_{jk}}}{\hat{\sigma}^2_{a_{jk}}}\right),$$

in which

$$\hat{\mu}_{a_{jk}} = \frac{\mathbf{x}'_{\mathbf{jk}}\left(\mathbf{y_k}-\hat{\boldsymbol{\mu}}_\mathbf{k}-\mathbf{X_{1k:jk-1}}\hat{\mathbf{a}}^{\mathbf{l+1}}_{\mathbf{1k:jk-1}}-\mathbf{X_{jk+1:mk}}\hat{\mathbf{a}}^{\mathbf{l}}_{\mathbf{jk+1:mk}}\right)}{\mathbf{x}'_{\mathbf{jk}}\mathbf{x_{jk}}+\sigma^2_{ek}/\sigma^2_{ak}},$$

and

$$\hat{\sigma}^2_{a_{jk}} = \frac{\sigma^2_{ek}}{\mathbf{x}'_{\mathbf{jk}}\mathbf{x_{jk}}+\sigma^2_{ek}/\sigma^2_{ak}}$$

(see Appendix for details).

b. For k=1,⋯c, sample a_{jk} from

$$f\left(a_{jk}|\gamma_j, \boldsymbol{\theta}_{j^-}, \mathbf{y}\right) = \begin{cases}\delta_0\left(a_{jk}\right) & \mathit{if}\ \gamma_j = 0\\ N\left(\hat{\mu}_{a_{jk}}, \hat{\sigma}^2_{a_{jk}}\right) & \mathit{if}\ \gamma_j = 1\end{cases}$$

Step 3. Sample w from Beta distribution

$f(w|\boldsymbol{\gamma}) = \frac{1}{B(\alpha,\beta)}w^{\alpha-1}(1-w)^{\beta-1}$, where $\alpha = m - \sum\gamma_j$ and $\beta = \sum\gamma_j$.

Step 4. For k=1,⋯, c, sample σ^2_{ak} from scaled inverse Chi-square distribution

$$\chi^{-2}\left(\nu_{ak}+\sum\gamma_j, \frac{\nu_{ak}s^2_{ak}+\sum\gamma_j a^2_{jk}}{\nu_{ak}+\sum\gamma_j}\right)$$

Step 5. For k=1,⋯, c, sample σ_{ek}^2 from scaled inverse Chi-square distribution

$$\chi^{-2}\left(\nu_{ek}+n_k, \frac{\nu_{ek}s_{ek}^2+\left(y_k-\mu_k-X_k a_k\right)'\left(y_k-\mu_k-X_k a_k\right)}{\nu_{ek}+n_k}\right)$$

Step 6. For k=1,⋯, c,

Sample μ_k from $N\left(\frac{\left(\mathbf{y}_k-\mathbf{X}_k\hat{\mathbf{a}}_k\right)'\left(\mathbf{y}_k-\mathbf{X}_k\hat{\mathbf{a}}_k\right)}{n_k}, \frac{\hat{\sigma}_{e_k}^2}{n_k}\right)$

Repeat Step 2 to 6 until a set number of iterations are reached.

It can be shown in step 2 that information from all populations are used to generate the latent variable γ_j, and the SNP effect a_{jk} is generated for each population.

Computer program

Computer programs were written in C language to implement the multi-task Bayesian learning model, single-task Bayesian learning and the simple data pooling method. Programs and source codes are available upon request.

Monte Carlo simulations

The aim of the simulation was to evaluate the performance of the proposed multi-task Bayesian learning model and to compare with the single-task model and the simple data pooling method. Different scenarios were considered that differ in number of QTL affecting the trait, correlations of the QTL effects between different populations, and the density of SNP panels used for genomic prediction.

Real genotypes from 458 Ayrshire and 2,298 Holstein bulls were used for simulations. All Ayrshire animals and 690 Holstein animals were genotyped on the Illumina BovineHD BeadChip (800 k) SNP panel, and the remaining 1,608 Holstein animals were genotyped on the Illumina BovineSNP50 BeadChip (50 k) SNP panel. SNPs meeting one of the following criteria were excluded: minor allele frequency (MAF) lower than 0.05, missing genotype rate greater than 0.10, highly correlated with any other SNP genotype (95% genotypes from two loci identical or in complementary). SNP locations were determined against Bovine genome assembly UMD3.1 [22]. SNPs with unknown locations or on sex chromosomes were discarded. SNPs filtered in one breed were also removed from the other breed. After editing, 246,668 SNPs from the 800 k panel and 28,206 SNPs from the 50 k panel were kept for analyses.

For ease of computation, only SNPs from the first 10 chromosomes were used. Four scenarios were considered with combinations of QTL number being either 20 or 200 and the correlation of QTL effects between the two populations being low (ρ = 0.2) or high (ρ = 0.8). For each scenario, QTL were randomly sampled from the 50 k panel. For each QTL j, allele substitution effects in the two populations, α_{j1} and α_{j2}, were sampled from a bivariate normal distribution with mean 0 and variance-covariance structure $\sum=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\sigma^2$ in which σ^2 = 1. Breeding value of each animal i, was calculated as $a_i=\sum_j X_j\alpha_j$ where X_j is the QTL genotypes coded as 0, 1 or 2, number of copies on an arbitrarily chosen allele. Total additive genetic variance were calculated within each breed as $\sigma_a^2=\sum_j 2p_j\left(1-p_j\right)\alpha_j^2$, where p_j is the QTL allele frequency. For each animal i, a residual effect e_i, was sampled from a normal distribution with mean 0 and variance σ_e^2, so that phenotypic value of the animal $y_i=a_i+e_i$. σ_e^2 was equal to σ_a^2 to give a trait with heritability of 0.5. Each scenario was replicated 10 times.

Four SNP panels of different density, the original 800 k panel, and the mimicked 400 k, 200 k, 100 k by selecting every 2nd, 4th, and 8th SNP, respectively, from the 800 k panel were used for genomic prediction. For each scenario described above, the simulated QTL were removed from the SNP panels. A special scenario was designed to keep all QTL genotypes in the 800 k panel for genomic prediction. The 50 k genotypes of 1,608 Holstein animals were imputed to 800 k using genotype imputation software FImpute developed by Sargolzaei et al. [23]. Imputation accuracy was evaluated on a set of 126 animals that have been genotyped on both the 50 k and 800 k panel, and the ratio of the genotypes that were correctly imputed was 0.9930.

For training and validation purposes, 393 Ayrshire and 2,084 Holstein animals born before 2004 were used as training set, 65 Ayrshire and 214 Holstein animals born in and after 2004 were used for validation. The number of animals used for genomic prediction is shown in Table 1. The simulated phenotypic values for training animals were used to derive SNP effects using different models. Degree of freedom of the inverse chi-square distributions for variances of SNP effects and residual effects were set to 4 and 10, respectively. The scale parameter S_a^2 was derived from the expected value of a

Table 1 Number of animals used for genomic prediction

	Ayrshire	Holstein
Training set	393	2084
Validation set	65	214
Total	458	2298

scaled inverse chi-square distributed random variable, i.e., $E(\sigma^2) = S^2\nu/(\nu - 2)$ and hence $S_a^2 = \sigma_a^2(\nu_a - 2)/\nu_a$ where σ_a^2 is the true additive genetic variance. S_e^2 was derived similarly. The Gibbs chain was run for 50,000 cycles with the first 10,000 discarded as burn-in. Marker effects were estimated as averages from all post burn-in samples. Genomic breeding values for animals in the validation set were estimated as the sum of population mean and all marker effects. Accuracy was evaluated as Pearson's correlation coefficient between genomic estimated breeding values and true breeding values for validation animals, i.e. r (GEBV, TBV). Regression of true breeding value on genomic estimated breeding values, b(TBV, GEBV), was also calculated to evaluate the bias of the genomic estimated breeding values.

Real data

Five milk production traits including milk yield, fat yield, protein yield, fat percentage and protein percentage were used for the same set of animals in the simulation study. Bull proofs (estimated breeding values from progeny testing, EBV) from April 2008 were used as phenotypes for training animals, and bull proofs from December 2011 were used for validation animals. All proofs were provided by Canadian Dairy Network (CDN) and had reliability above 0.65.

Two SNP panels with 28,206 SNPs from the 50 k panel and 246,668 SNPs from the 800 k panel were used for genomic prediction. Degree of freedom of the inverse chi-square distributions for variances of SNP effects and residual effects were set to 4 and 10, respectively. Scale parameters for the two distributions were derived in the same way as described in the simulation study, but instead of using true additive genetic variance and residual variance, estimated variances were used. Estimated additive genetic variances and residual variance were obtained from ASReml [24] by fitting an animal model with a population mean, animal effect, and random residual effect. The pedigree contained 15,731 animals for the Holstein population and 4,926 animals for Ayrshire. For 28,206 SNPs, the Gibbs sampling procedure was run for 50,000 iterations with the first 10,000 discarded as burn-in; and for 246,668 SNPs, the Gibbs sampling procedure was run for 100,000 iterations with the first 50,000 discarded as burn-in. Burn-in period was determined by visually inspecting the Gibbs chain. All samples after the burn-in period were kept. SNP effects were estimated by averaging all samples after the burn-in period. After the estimation of SNP effects, the GEBV was calculated for animals in the validation set by summing up the population mean and all the SNP effects. Accuracy was measured as Pearson's correlation coefficient between GEBV and the 2011 bull proofs for validation animals.

Results

Monte Carlo simulation study

Table 2 shows the accuracy of genomic prediction with simulated 20 QTL. In the Ayrshire validation set, multi-task Bayesian learning model performed the best among the three methods within each SNP panel used under the scenario with either a low (ρ = 0.2) or high (ρ = 0.8) correlation of simulated QTL effects between Ayrshire and Holstein populations. The greatest increase of accuracy was observed when ρ was 0.8 and when density of the SNP panel was the highest (accuracy increased from 0.67 for the single-task method to 0.83 for the multi-task method). Simply pooling data together substantially reduced the prediction accuracy in Ayshire when ρ was 0.2. The greatest reduction of accuracy was observed when ρ was 0.2 and when density of the SNP panel was the highest (accuracy decreased from 0.71 for the single-task method to 0.56 for the pooling method). When ρ was 0.8, the pooling method had an increased accuracy in Ayrshire compared with the single-task method. The pooling method produced substantially lower accuracy than the multi-task model in the Ayrshire validation set, especially when QTL effects were lower correlated between the two populations. In the Holstein validation set, the multi-task model performed similar or slightly better than the single-task model. The pooling method had a slightly reduction of accuracy when ρ was 0.2, and had a similar accuracy compared with the single-task method when ρ was 0.8. Within each method, increasing density of the SNP panel generally improved prediction accuracy in both the Ayrshire and Holstein populations. Table 2 also shows the slopes of regression of true breeding values on the GEBV. Regression coefficients of true breeding values on GEBV were less than one for the pooling method indicating that the GEBV were inflated.

Table 3 shows the accuracy of genomic prediction with simulated 200 QTL. In the Ayrshire validation set, the multi-task model had slightly higher prediction accuracy within each SNP panel compared with the single-task model under either scenario with low (ρ =0.2) or high (ρ =0.8) correlated QTL effects between the Ayrshire and Holstein populations. Pooling method performed similar or slightly worse compared with single-task model when ρ was 0.2, and generally performed better when ρ was 0.8. The pooling method also performed slightly better than the multi-task model when ρ was 0.8 and when density of the SNP panel was relatively high. When ρ was 0.2, regression coefficients of true breeding values on GEBV were less than one for the pooling method indicating that the GEBV were inflated. The three methods performed similar in the Holstein validation set. Overall, the accuracy was lower compared with scenarios when only

Table 2 Accuracy {expressed as correlations between true breeding values (TBV) and genomic estimated breeding values [GEBV; r(TBV, GEBV)]}, and slopes [b(TBV, GEBV)] of regression of TBV on GEBV for genomic prediction with 20 simulated QTL

	Ayrshire			**Holstein**		
SNP panel	**Single-task**	**Pooling**	**Multi-task**	**Single-task**	**Pooling**	**Multi-task**
r(TBV, GEBV)						
ρ = 0.2						
800 k	0.71 ± 0.02	0.56 ± 0.04	0.75 ± 0.02	0.91 ± 0.01	0.90 ± 0.01	0.91 ± 0.01
400 k	0.64 ± 0.04	0.53 ± 0.04	0.73 ± 0.03	0.90 ± 0.01	0.86 ± 0.01	0.90 ± 0.01
200 k	0.60 ± 0.05	0.50 ± 0.04	0.68 ± 0.02	0.88 ± 0.01	0.84 ± 0.01	0.88 ± 0.01
100 k	0.57 ± 0.04	0.47 ± 0.04	0.63 ± 0.03	0.84 ± 0.01	0.81 ± 0.02	0.84 ± 0.01
ρ = 0.8						
800 k	0.67 ± 0.05	0.76 ± 0.02	0.83 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.93 ± 0.01
400 k	0.66 ± 0.05	0.72 ± 0.02	0.80 ± 0.01	0.90 ± 0.01	0.89 ± 0.01	0.90 ± 0.01
200 k	0.66 ± 0.04	0.68 ± 0.02	0.76 ± 0.01	0.86 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
100 k	0.61 ± 0.04	0.63 ± 0.04	0.72 ± 0.02	0.83 ± 0.01	0.83 ± 0.01	0.84 ± 0.01
b(TBV, GEBV)						
ρ = 0.2						
800 k	1.06 ± 0.05	0.73 ± 0.05	1.04 ± 0.05	0.98 ± 0.02	1.01 ± 0.01	0.98 ± 0.01
400 k	1.06 ± 0.05	0.76 ± 0.06	1.06 ± 0.05	0.99 ± 0.02	1.00 ± 0.01	0.98 ± 0.01
200 k	1.02 ± 0.06	0.75 ± 0.05	1.03 ± 0.04	1.00 ± 0.01	0.99 ± 0.01	0.98 ± 0.01
100 k	1.00 ± 0.06	0.75 ± 0.05	1.03 ± 0.06	1.00 ± 0.01	1.00 ± 0.02	0.99 ± 0.01
ρ = 0.8						
800 k	1.10 ± 0.06	0.90 ± 0.03	1.10 ± 0.04	0.99 ± 0.02	1.00 ± 0.02	0.99 ± 0.02
400 k	1.09 ± 0.06	0.89 ± 0.04	1.07 ± 0.04	0.99 ± 0.02	0.99 ± 0.01	0.99 ± 0.02
200 k	1.14 ± 0.06	0.89 ± 0.06	1.04 ± 0.05	0.98 ± 0.02	1.00 ± 0.02	0.98 ± 0.03
100 k	1.08 ± 0.08	0.85 ± 0.06	1.06 ± 0.05	0.98 ± 0.03	0.99 ± 0.03	0.98 ± 0.03

20 QTL were simulated regardless of the methods used.

To evaluate the performance of the three methods under situations where all QTL genotypes can be acquired and included for genomic prediction, Table 4 shows the accuracy of genomic prediction when QTL genotypes are included together with SNP marker genotypes for training and validation. When 20 QTL were simulated, using multi-task model improved the accuracy by 0.16 and 0.22 in the Ayrshire validation set for scenarios where correlation of QTL effects between the Ayrshire and Holstein populations was 0.2 and 0.8, respectively. Using the pooling method reduced the accuracy in Ayrshire by 0.12 when ρ was 0.2, but increased the accuracy by 0.17 when ρ was 0.8. When 200 QTL were simulated, using the multi-task model increased the prediction accuracy by 0.03 and 0.07, respectively, for ρ equal to 0.2 and 0.8. The pooling method reduced the accuracy by 0.04 when ρ was 0.2, and increased the accuracy by 0.11 when ρ was 0.8. The pooling method outperformed the multi-task model only for the scenario where 200 QTL were simulated with their effects highly correlated between the two populations. In the Holstein validation set, the multi-task model had similar or slightly higher accuracy compared with the single-task model. The pooling method had similar accuracy as the single-task model when ρ was 0.8. When ρ was 0.2, the pooling method had slightly lower accuracy compared with the single-task model. Regression coefficients of true breeding values on GEBV were less than one for the pooling method except for the scenario of 200 simulated QTL with effects highly correlated between the two populations, indicating that the GEBV predicted by the pooling method were inflated.

Real data analysis

Table 5 shows the accuracy of genomic prediction for milk production traits using real data. For the Ayrshire validation set, the multi-task model increased the accuracy by up to 0.07 compared with the single-task model, while the simple data pooling method resulted in reduced accuracy in general. The greatest increase in accuracy using

Table 3 Accuracy {expressed as correlations between true breeding values (TBV) and genomic estimated breeding values [GEBV; r(TBV, GEBV)]}, and slopes [b(TBV, GEBV)] of regression of TBV on GEBV for genomic prediction with 200 simulated QTL

	Ayrshire			Holstein		
SNP panel	**Single-task**	**Pooling**	**Multi-task**	**Single-task**	**Pooling**	**Multi-task**
r(TBV, GEBV)						
ρ = 0.2						
800 k	0.46 ± 0.02	0.44 ± 0.04	0.47 ± 0.03	0.77 ± 0.01	0.76 ± 0.01	0.77 ± 0.01
400 k	0.46 ± 0.02	0.43 ± 0.03	0.47 ± 0.02	0.76 ± 0.01	0.76 ± 0.01	0.76 ± 0.01
200 k	0.46 ± 0.02	0.42 ± 0.04	0.47 ± 0.02	0.75 ± 0.01	0.75 ± 0.01	0.75 ± 0.01
100 k	0.45 ± 0.02	0.41 ± 0.03	0.46 ± 0.03	0.74 ± 0.01	0.74 ± 0.01	0.74 ± 0.01
ρ = 0.8						
800 k	0.54 ± 0.04	0.57 ± 0.02	0.56 ± 0.03	0.74 ± 0.02	0.75 ± 0.01	0.75 ± 0.02
400 k	0.54 ± 0.03	0.56 ± 0.03	0.55 ± 0.03	0.74 ± 0.02	0.74 ± 0.02	0.74 ± 0.02
200 k	0.54 ± 0.03	0.56 ± 0.02	0.55 ± 0.03	0.73 ± 0.02	0.73 ± 0.02	0.73 ± 0.02
100 k	0.53 ± 0.03	0.52 ± 0.02	0.54 ± 0.03	0.72 ± 0.02	0.72 ± 0.02	0.72 ± 0.02
b(TBV, GEBV)						
ρ = 0.2						
800 k	1.14 ± 0.08	0.89 ± 0.09	1.16 ± 0.09	1.07 ± 0.01	1.09 ± 0.01	1.07 ± 0.01
400 k	1.14 ± 0.09	0.91 ± 0.08	1.13 ± 0.09	1.07 ± 0.01	1.09 ± 0.01	1.07 ± 0.01
200 k	1.14 ± 0.09	0.88 ± 0.08	1.14 ± 0.09	1.08 ± 0.01	1.09 ± 0.01	1.08 ± 0.01
100 k	1.15 ± 0.09	0.89 ± 0.09	1.15 ± 0.09	1.09 ± 0.02	1.11 ± 0.03	1.09 ± 0.02
ρ = 0.8						
800 k	1.12 ± 0.11	1.00 ± 0.06	1.07 ± 0.09	1.01 ± 0.02	1.00 ± 0.02	1.01 ± 0.02
400 k	1.11 ± 0.11	1.02 ± 0.07	1.08 ± 0.09	1.01 ± 0.02	1.00 ± 0.02	1.01 ± 0.02
200 k	1.12 ± 0.11	1.04 ± 0.07	1.11 ± 0.09	1.01 ± 0.02	1.01 ± 0.02	1.01 ± 0.02
100 k	1.11 ± 0.11	0.99 ± 0.07	1.10 ± 0.10	1.01 ± 0.02	1.01 ± 0.02	1.01 ± 0.02

Table 4 Accuracy {expressed as correlations between true breeding values (TBV) and genomic estimated breeding values [GEBV; r(TBV, GEBV)]}, and slopes [b(TBV, GEBV)] of regression of TBV on GEBV for genomic prediction with simulated QTL genotypes included for training and validation

		Ayrshire			Holstein		
No. of QTL	**ρ**	**Single-task**	**Pooling**	**Multi-task**	**Single-task**	**Pooling**	**Multi-task**
r(TBV, GEBV)							
20	0.2	0.76 ± 0.04	0.64 ± 0.04	0.92 ± 0.01	0.96 ± 0.01	0.93 ± 0.01	0.97 ± 0.01
20	0.8	0.71 ± 0.05	0.88 ± 0.02	0.93 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.97 ± 0.01
200	0.2	0.57 ± 0.03	0.53 ± 0.03	0.60 ± 0.04	0.77 ± 0.01	0.76 ± 0.01	0.78 ± 0.01
200	0.8	0.54 ± 0.04	0.65 ± 0.02	0.61 ± 0.04	0.76 ± 0.02	0.78 ± 0.02	0.77 ± 0.01
b(TBV, GEBV)							
20	0.2	1.01 ± 0.06	0.83 ± 0.08	0.99 ± 0.03	1.00 ± 0.01	1.00 ± 0.02	1.00 ± 0.01
20	0.8	1.04 ± 0.08	0.89 ± 0.06	0.97 ± 0.03	0.98 ± 0.01	1.01 ± 0.02	0.99 ± 0.01
200	0.2	1.23 ± 0.10	0.89 ± 0.06	1.16 ± 0.08	1.00 ± 0.03	1.03 ± 0.03	1.01 ± 0.03
200	0.8	1.10 ± 0.09	1.06 ± 0.06	1.12 ± 0.09	0.98 ± 0.03	0.97 ± 0.03	0.98 ± 0.02

Table 5 Accuracy of genomic prediction of breeding values for milk production traits

	Ayrshire			Holstein		
Trait	Single-task	Pooling	Multi-task	Single-task	Pooling	Multi-task
No. of SNP: 28,206						
Milk yield	0.52	0.44	0.54	0.66	0.65	0.66
Fat yield	0.64	0.55	0.66	0.63	0.64	0.63
Protein yield	0.70	0.60	0.70	0.69	0.69	0.68
Fat %	0.66	0.58	0.72	0.74	0.74	0.74
Protein %	0.48	0.51	0.55	0.67	0.68	0.67
No. of SNP: 246,668						
Milk yield	0.54	0.53	0.55	0.64	0.64	0.64
Fat yield	0.67	0.62	0.67	0.63	0.64	0.63
Protein yield	0.72	0.68	0.72	0.66	0.66	0.66
Fat %	0.66	0.65	0.69	0.77	0.77	0.78
Protein %	0.51	0.42	0.53	0.71	0.68	0.70

multi-task model compared with single-task model was for fat percentage (0.06) and protein percentage (0.07) when 28,206 SNPs were used. For the Holstein validation set, the single-task model, simple data pooling method, and the multi-task model performed similar regardless of the traits studied.

Discussion

Traditionally, genomic prediction with data from multiple populations were implemented either by running genomic prediction within each population (single-task) or by simply pooling data together. Single-task genomic prediction cannot utilize information from other populations and therefore, the accuracy of genomic prediction is largely determined by the size of training data set [2,3]. For breeds with only a small number of animals having both DNA marker genotypes and phenotype data, the accuracy of genomic prediction can be low [25]. Combining the data with other breed populations has the potential to improve the prediction accuracy. It is however, difficult to effectively account for the differences of SNP effects among different populations by simply pooling data together. If the marker density is low or the populations are divergent from each other, simply pooling data together may result in unfavorable prediction accuracies [6]. The multi-task Bayesian learning model proposed in this study uses information from all populations simultaneously while allowing the SNP effects to vary in different populations. Different populations share information through a common set of latent indicator variables. When the target trait has a similar genetic background in related populations, it is reasonable to assume that some shared QTL affecting a common trait in different populations. However, the linkage disequilibrium phase between SNP markers and QTL are likely to be inconsistent, especially when the marker density is low. Therefore, the multi-task Bayesian learning model is more flexible about the SNP effects and is likely to have better performance than a simple data pooling method.

Results from simulation studies support the use of multi-task Bayesian model for multi-population genomic prediction especially when there are a few QTL affecting the trait. For the scenario where a few QTL affect the trait, the increase of accuracy by using multi-task model was greater when QTL effects had a higher correlation between two populations. The accuracy was further increased when QTL genotypes were included for training and validation. These results are expected as the higher the correlation of QTL effects between two populations, the more informative information the two populations can share. Including QTL genotypes also increased the amount of information to be shared between the two populations. Results suggest that the proposed multi-task Bayesian learning model is effective in combining information from multi-populations to improve accuracy of genomic prediction.

Results from simulation also showed that simply pooling data together may reduce the accuracy of genomic prediction if QTL effects were lowly correlated between two populations or if a relatively low density SNP panel was used. When QTL effects had a high correlation between two populations and the SNP panel was high, the pooling method increased the accuracy compared with single-task model. When the correlation of QTL effects between two populations was high, the pooling method was inferior to the multi-task model if the number of QTL was small, but outperformed the multi-task model if the number of QTL was large. These results are reasonable since the pooling method assumes the same SNP effects among different populations. This assumption will be violated if the correlation of QTL effects is low between two populations and may result in poor performance of the pooling method. The multi-task model proposes to share information through the same latent indicator variables. Correlations of QTL effects among different populations are not explicitly used and therefore, information across populations may not be fully utilized if such correlations are high. This might be why the pooling method still outperformed the multi-task model when the number of QTL is large and the correlation of QTL effects between two populations was high.

Results from real data analyses in this study showed that the multi-task Bayesian learning model produced a similar or higher accuracy compared to the single-task model, and that simply pooling data together resulted in a reduced accuracy when the marker density was low. SNP effects could be different across populations, especially when lower density markers were used [6]. These results are in agreement with results from simulation studies.

Gains of accuracy by using the multi-task model were higher for fat percentage and protein percentage traits than for other traits. This is likely due to that large QTL or genes such as DGAT1, have larger influence on the percentage traits than on the yield traits [26,27]. These results agreed with simulation studies which showed that the multi-task model performed better when there are fewer QTL affecting the trait.

In this study, only two dairy cattle breeds were considered, and the multi-task Bayesian learning model was shown to be effective and more beneficial to the population with a smaller data size. For the larger population, using multi-task model did not produce much improvement. The little improvement in the larger population could be due to that the smaller population is too small to be able to have a significant impact on the larger population. In practice, there are situations where many populations are to be combined for analysis with each one contributing a small amount of data, a typical example as in beef cattle production. It would be interesting to test the performance of the multi-task model under such scenarios.

The current multi-task model considered additive genetic effects only. Non-additive genetic effects, which have gained growing interests with availability of genomics information [28-30], can also be accommodated into the multi-task model. A similar strategy used in this study to sharing information across populations for additive genetic effects may also be applied to non-additive genetic effects. Inclusion of non-additive genetic effects in the multi-task model warrants further investigation.

The proposed multi-task Bayesian learning model used a spike and slab mixture distribution to conduct variable selection and shrinkage for SNP effects. This prior setting is similar to that in the BayesCπ method proposed by Habier et al. [21]. Other types of mixture distributions currently being used for genomic prediction [1,19,20], can also be adapted to the multi-task model. The strategy used in the multi-task Bayesian learning model allows different populations to share information through a common latent variable assumed for each SNP. Such strategy has been shown effective in both simulation and real data studies; however, other strategies may also be exploited, for example, by modeling the joint distribution of the SNP effects among different populations. Further investigations are required to evaluate alternative strategies for sharing information across populations.

In this study, the Bayesian models were implemented via a Gibbs sampling algorithm adopting the residual-update computing strategy proposed by Legarra and Misztal [31]. Computing time for this algorithm is proportional to the number of animals and number of SNPs in the model. For the training sample size of 2,477 animals and genotypes of 246,668 SNPs, the proposed multi-task model requires 44 CPU hours to complete 150,000 Gibbs sampling cycles on a Linux cluster system with Intel X5675 3.07GHz CPU. With increased training sample size and increasing marker density to a very high density panel or even sequence data, computing burdens would become a concern. A recent study [32] has improved the residual-update algorithm resulting in the CPU time reduced by 35.3 to 43.3%. The authors in the same study [32] also proposed an alternative algorithm which reduced CPU time by 74.5 to 93.0%. Approximation algorithms based on expectation-maximization (EM) [33-35] and variational Bayes [36] have also been proposed to replace the time consuming Markov chain Monte Carlo (MCMC) sampling based algorithms. Adapting these algorithms to accommodate the multi-task Bayesian learning model would be of great interest.

Conclusions

A multi-task Bayesian learning model was proposed for multi-population genomic prediction. The multi-task model shares information across populations through a common set of latent indicator variables while allowing the SNP effects to vary in different populations. Simulation studies and real data analysis suggest that the proposed multi-task Bayesian learning model is effective and beneficial to populations where a small number of training animals are available. Accuracy of genomic prediction in small populations can be improved by using the multi-task model especially for traits affected by a few QTL with large effects.

Appendix

Sampling of r_j in the multi-task Bayesian learning model

With the assumption of independence between θ_{j^-} and r_j, by Bayes' theorem, one has

$$f(r_j=1|\theta_{j^-},y)=\frac{f(\mathbf{y}|r_j=1,\boldsymbol{\theta}_{\mathbf{j}^-})f(r_j=1)}{f(\mathbf{y}|r_j=1,\boldsymbol{\theta}_{\mathbf{j}^-})f(r_j=1)+f(\mathbf{y}|r_j=0,\boldsymbol{\theta}_{\mathbf{j}^-})f(r_j=0)}$$
$$=\frac{1}{1+q_j}, \tag{A.1}$$

where

$$q_j=\frac{f(\mathbf{y}|r_j=0,\boldsymbol{\theta}_{\mathbf{j}^-})f(r_j=0)}{f(\mathbf{y}|r_j=1,\boldsymbol{\theta}_{\mathbf{j}^-})f(r_j=1)} \tag{A.2}$$

$$=\frac{f(r_j=0)}{f(r_j=1)}\prod_{k=1}^{c}\frac{f(\mathbf{y_k}|r_j=0,\boldsymbol{\theta}_{\mathbf{j}^-\mathbf{k}})}{f(\mathbf{y_k}|r_j=1,\boldsymbol{\theta}_{\mathbf{j}^-\mathbf{k}})}.$$

Similarly,

$$f(r_j=0|\boldsymbol{\theta}_{\mathbf{j}^-},\mathbf{y})=\frac{q_j}{1+q_j}.$$

Next, the conditional likelihoods $f(\mathbf{y_k}|r_j=0,\boldsymbol{\theta}_{\mathbf{j}^-\mathbf{k}})$ and $f(\mathbf{y_k}|r_j=1,\boldsymbol{\theta}_{\mathbf{j}^-\mathbf{k}})$ will be derived.

Suppose one is at the $(l+1)^{th}$ Gibbs sampling iteration, and wants to sample γ_j and a_{jk}, the linear regression model given all parameters except γ_j and a_{jk} can be written as:

$$\mathbf{y}_k^*=\mathbf{x}_{jk}a_{jk}+\mathbf{e}_k,$$

Where $\mathbf{y}_k^*=\mathbf{y}_k-\hat{\mu}_k-\mathbf{X}_{1k:jk-1}\hat{\mathbf{a}}_{1k:jk-1}^{l+1}-\mathbf{X}_{jk+1:mk}\hat{\mathbf{a}}_{jk+1:mk}^{l}$.

Given the priors that $(a_{jk}|\gamma_j)\sim\gamma_jN(0,\sigma_{ak}^2)+\left(1-\gamma_j\right)\delta_0(a_{jk})$ and $(\mathbf{e_k}|\sigma_{ek}^2)\sim N(\mathbf{0},\mathbf{I}\sigma_{ek}^2)$, the conditional likelihood of $\mathbf{y_k}$ can be written as:

$$f(\mathbf{y_k}|r_j=0,\boldsymbol{\theta}_{\mathbf{j}^-\mathbf{k}})=\left(2\pi\sigma_{ek}^2\right)^{-n_k/2}\exp\left(-\frac{\mathbf{y_k^*}'\mathbf{y_k^*}}{2\sigma_{ek}^2}\right) \tag{A.3}$$

And

$$f(\mathbf{y_k}|r_j=1,\boldsymbol{\theta}_{\mathbf{j}^-\mathbf{k}})=(2\pi)^{-n_k/2}\left|\mathbf{V}_{\mathbf{y_k^*}}\right|^{-1/2}\exp\left(-\frac{\mathbf{y_k^*}'\mathbf{V}_{\mathbf{y_k^*}}^{\mathbf{-1}}\mathbf{y}^*}{2}\right) \tag{A.4}$$

Where $V_{y_k^*}=\mathbf{x_{jk}}\mathbf{x_{jk}}'\sigma_{ak}^2+I\sigma_{ek}^2$ is the (co-) variance matrix of $\mathbf{y_k^*}$ given that r_j=1.

Then, $\left|\mathbf{V}_{y_k^*}\right|$ and $\mathbf{V}_{y_k^*}^{-1}$ are derived.

$$\left|\mathbf{V}_{y_k^*}\right|=\left|\mathbf{x}_{jk}\mathbf{x}_{jk}'\sigma_{ak}^2+\mathbf{I}\sigma_{ek}^2\right|=\sigma_{ek}^{2n_k}\left|\mathbf{x}_{jk}\frac{\sigma_{ak}}{\sigma_{ek}}\mathbf{x}_{jk}'\frac{\sigma_{ak}}{\sigma_{ek}}+\mathbf{I}\right|,$$

By applying Sylvester's determinant theorem, one has

$$\left|\mathbf{V}_{\mathbf{y_k^*}}\right|=\sigma_{ek}^{2n_k}\left(\mathbf{x_{jk}'}\mathbf{x_{jk}}\frac{\sigma_{ak}^2}{\sigma_{ek}^2}+1\right) \tag{A.5}$$

It can be easily verify that $V_{y^*}^{-1}$ is:

$$\mathbf{V}_{\mathbf{y_k^*}}^{\mathbf{-1}}=\sigma_{ek}^{-2}\left(\mathbf{I}-\frac{\mathbf{x_{jk}}\mathbf{x_{jk}'}}{\mathbf{x_{jk}'}\mathbf{x_{jk}}+\sigma_{ek}^2/\sigma_{ak}^2}\right) \tag{A.6}$$

Substituting A.5 and A.6 into A.4, one has

$$f(\mathbf{y}|r_j=1,\boldsymbol{\theta}_{\mathbf{j}^-})=\left(2\pi\sigma_{ek}^2\right)^{-n_k/2}\left(\mathbf{x_{jk}'}\mathbf{x_{jk}}\frac{\sigma_{ak}^2}{\sigma_{ek}^2}+1\right)^{-1/2}$$
$$\exp\left\{-\frac{\mathbf{y_k^*}'\mathbf{y_k^*}-\left(\mathbf{x_{jk}'}\mathbf{y_k^*}\right)^2/\left(\mathbf{x_{jk}'}\mathbf{x_{jk}}+\sigma_{ek}^2/\sigma_{ak}^2\right)}{2\sigma_{ek}^2}\right\} \tag{A.7}$$

Denote $\hat{\mu}_{a_{jk}}=\frac{\mathbf{x_{jk}'}\left(\mathbf{y_k}-\mathbf{X_{1k:jk-1}}\hat{\mathbf{a}}_{\mathbf{1k:jk-1}}^{\mathbf{l+1}}-\mathbf{X_{jk+1:mk}}\hat{\mathbf{a}}_{\mathbf{jk+1:mk}}^{\mathbf{l}}\right)}{\mathbf{x_{jk}'}\mathbf{x_{jk}}+\sigma_{ek}^2/\sigma_{ak}^2}$, $\hat{\sigma}_{a_{jk}}^2=\frac{\sigma_{ek}^2}{\mathbf{x_{jk}'}\mathbf{x_{jk}}+\sigma_{ek}^2/\sigma_{ak}^2}$, and substitute (A.7) and (A.3) into (A.2), one gets

$$q_j=\frac{w}{1-w}\sqrt{\prod_k\left(\mathbf{x_{jk}'}\mathbf{x_{jk}}\frac{\sigma_{ak}^2}{\sigma_{ek}^2}+1\right)}\exp\left(-\frac{1}{2}\sum_k\frac{\hat{\mu}_{a_{jk}}^2}{\hat{\sigma}_{a_{jk}}^2}\right) \tag{A.8}$$

Finally, γ_j can be drawn from a Bernoulli distribution with probability 1/(1+q$_j$).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LC, CL and FS contributed to the design of the study. LC, CL, SM and FS contributed to the development of the methods and discussions of analysis and results. LC performed the data analysis and drafted the manuscript. All authors read, provided comments and agreed on the contents of the manuscript.

Acknowledgements

The authors thank the financial supports from the Natural Sciences and Engineering Research Council of Canada, L'Alliance Boviteq Inc., the AAFC Livestock Genetics & Genomics Program funds, and the AAFC A-base peer-reviewed research project (RBPI-1139). Canadian dairy network was acknowledged for providing data for this study. The authors also thank Drs Mehdi Sargozaei and Niel Karrow from University of Guelph, Dr Yuanming Zhang from journal of BMC Genetics and the other anonymous reviewer for their critical review and comments on the manuscript. This research has been enabled by the use of computing resources provided by WestGrid and Compute/Calcul Canada.

Author details

[1]Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G 2P5, Canada. [2]Department of Animal and Poultry Science, University of Guelph, Guelph, ON N1G 2W1, Canada. [3]Agriculture and Agri-Food Canada, Lacombe Research Centre, 6000 C&E Trail, Lacombe, AB T4L 1W1, Canada.

Received: 7 August 2013 Accepted: 28 April 2014
Published: 3 May 2014

References

1. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**(4):1819–1829.
2. Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136**(2):245–257.
3. Hayes BJ, Goddard ME: **Technical note: prediction of breeding values using marker-derived relationship matrices.** *J Anim Sci* 2008, **86**(9):2089–2092.
4. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: genomic selection in dairy cattle: progress and challenges (vol 92, pg 433, 2009).** *J Dairy Sci* 2009, **92**(3):1313–1313.
5. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited review: Reliability of genomic predictions for north american holstein bulls.** *J Dairy Sci* 2009, **92**(1):16–24.
6. de Roos APW, Hayes BJ, Goddard ME: **Reliability of genomic predictions across multiple populations.** *Genetics* 2009, **183**(4):1545–1553.
7. Hayes B, Bowman P, Chamberlain A, Verbyla K, Goddard M: **Accuracy of genomic breeding values in multi-breed dairy cattle populations.** *Genet Sel Evol* 2009, **41**(1):51.
8. Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Fuerst C, Emmerling R, Solkner J, Goddard ME, Hayes BJ: **Short communication: genomic selection using a multi-breed, across-country reference population.** *J Dairy Sci* 2011, **94**(5):2625–2630.
9. 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):4114–4129.
10. Brondum RF, Su GS, Lund MS, Bowman PJ, Goddard ME, Hayes BJ: **Genome position specific priors for genomic prediction.** *BMC Genomics* 2012, **13**(1):543.
11. Caruana R: **Multitask learning.** *Mach Learn* 1997, **28**(1):41–75.
12. Li X, Bilmes J: **Regularized adaptation of discriminative classifiers.** In *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 14-19 May 2006; Toulouse.* IEEE 2006(vol. 1):237-240.
13. Lu Y, Lu F, Sehgal S, Gupta S, Du J, Tham CH, Green P, Wan V: **Multitask Learning In Connectionist Speech Recognition**. In *Proceedings of the Tenth Australian International Conference on Speech Science & Technology: 8-10 December 2004; Sydney.* Edited by Cassidy S, Cox F, Mannell R, Palethorpe S. Canberra: Australian Speech Science and Technology Association Inc; 2004:312–315.
14. Yuan X-T, Yan S: **Visual classification with multi-task joint sparse representation.** In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 13-18 June 2010; San Francisco:* IEEE; 2010:3493–3500.
15. Jacob L, Vert J-P: **Efficient peptide–mhc-i binding prediction for alleles with few known binders.** *Bioinformatics* 2008, **24**(3):358–366.
16. Widmer C, Leiva J, Altun Y, Rätsch G: **Leveraging sequence classification by taxonomy-based multitask learning**. In *Research in Computational Molecular Biology.* Heidelberg: Springer Berlin; 2010:522–534.
17. Yang W-H, Dai D-Q, Yan H: **Finding correlated biclusters from gene expression data.** *Knowl Data Eng, IEEE T* 2011, **23**(4):568–584.
18. Puniyani K, Kim S, Xing EP: **Multi-population gwa mapping via multi-task regularized regression.** *Bioinformatics* 2010, **26**(12):i208–i216.
19. Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF: **Accuracy of genomic selection using different methods to define haplotypes.** *Genetics* 2008, **178**(1):553–561.
20. Verbyla KL, Hayes BJ, Bowman PJ, Goddard ME: **Accuracy of genomic selection using stochastic search variable selection in australian holstein friesian dairy cattle.** *Genet Res* 2009, **91**(5):307–311.
21. Habier D, Fernando RL, Kizilkaya K, Garrick DJ: **Extension of the bayesian alphabet for genomic selection.** *BMC Bioinformatics* 2011, **12**(1):186.
22. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS: **A whole-genome assembly of the domestic cow, bos taurus.** *Genome Biol* 2009, **10**(4):R42.
23. Sargolzaei M, Chesnais JP, Schenkel FS: **Fimpute - an efficient imputation algorithm for dairy cattle populations.** *J Dairy Sci* 2011, **94**(E-Suppl. 1):421.
24. Gilmour AR, Gogel BJ, Cullis BR, Thompson R: **Asreml user guide release 3.0**. In *Hemel Hempstead, HP1 1ES.* UK: VSN International Ltd; 2009.
25. Brito FV, Neto JB, Sargolzaei M, Cobuci JA, Schenkel FS: **Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle.** *BMC Genet* 2011, **12**(1):80.
26. Grisart B, Farnir F, Karim L, Cambisano N, Kim JJ, Kvasz A, Mni M, Simon P, Frere JM, Coppieters W, *et al*: **Genetic and functional confirmation of the causality of the dgat1 k232a quantitative trait nucleotide in affecting milk yield and composition.** *Proc Natl Acad Sci U S A* 2004, **101**(8):2398–2403.
27. Pimentel Eda C, Erbe M, Konig S, Simianer H: **Genome partitioning of genetic variation for milk production and composition traits in holstein cattle.** *Front Genet* 2011, **2**:19.
28. Su G, Christensen OF, Ostersen T, Henryon M, Lund MS: **Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers.** *PLoS One* 2012, **7**(9):e45293.
29. Toro MA, Varona L: **A note on mate allocation for dominance handling in genomic selection.** *Genet Sel Evol* 2010, **42**:33.
30. Vitezica ZG, Varona L, Legarra A: **On the additive and dominant variance and covariance of individuals within the genomic selection scope.** *Genetics* 2013, **195**(4):1223–1230.
31. Legarra A, Misztal I: **Technical note: computing strategies in genome-wide selection.** *J Dairy Sci* 2008, **91**(1):360–366.
32. Calus MP: **Right-hand-side updating for fast computing of genomic breeding values.** *Genet Sel Evol* 2014, **46**(1):24.
33. Hayashi T, Iwata H: **Em algorithm for bayesian estimation of genomic breeding values.** *BMC Genet* 2010, **11**:3.
34. 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**:1.
35. Shepherd RK, Meuwissen TH, Woolliams JA: **Genomic selection and complex trait prediction using a fast em algorithm applied to genome-wide markers.** *BMC Bioinformatics* 2010, **11**(1):529.
36. Li ZT, Sillanpaa MJ: **Estimation of quantitative trait locus effects with epistasis by variational bayes algorithms.** *Genetics* 2012, **190**(1):231–249.

doi:10.1186/1471-2156-15-53
Cite this article as: Chen *et al.*: Multi-population genomic prediction using a multi-task Bayesian learning model. *BMC Genetics* 2014 **15**:53.